

## **EXHIBIT 2**

(12) **United States Patent**  
**Cox**

(10) **Patent No.:** **US 8,205,237 B2**

(45) **Date of Patent:** \*Jun. 19, 2012

(54) **IDENTIFYING WORKS, USING A SUB-LINEAR TIME SEARCH, SUCH AS AN APPROXIMATE NEAREST NEIGHBOR SEARCH, FOR INITIATING A WORK-BASED ACTION, SUCH AS AN ACTION ON THE INTERNET**

(76) Inventor: **Ingemar J. Cox**, London (GB)

(\*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 594 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: 11/977,202

(22) Filed: **Oct. 23, 2007**

(65) **Prior Publication Data**

US 2008/0060036 A1 Mar. 6, 2008

### Related U.S. Application Data

(63) Continuation of application No. 11/445,928, filed on Jun. 2, 2006, which is a continuation-in-part of application No. 09/950,972, filed on Sep. 13, 2001, now Pat. No. 7,058,223.

(60) Provisional application No. 60/232,618, filed on Sep. 14, 2000.

(51) **Int. Cl.**  
**H04N 7/173** (2011.01)

(52) **U.S. Cl.** ..... 725/110

(58) **Field of Classification Search** ..... None  
See application file for complete search history.

(56) **References Cited**

## U.S. PATENT DOCUMENTS

3,919,479	A	11/1975	Moon et al.
4,230,990	A	10/1980	Lert, Jr. et al.
4,450,531	A	5/1984	Kenyon et al.
4,495,526	A	1/1985	Baranoff-Rossine

4,499,601 A	2/1985	Matthews
4,511,917 A	4/1985	Kohler et al.
4,547,804 A	10/1985	Greenberg
4,634,966 A	1/1987	Nakatani et al.
4,639,779 A	1/1987	Greenberg
4,677,455 A	6/1987	Okajima
4,677,466 A	6/1987	Lert, Jr. et al.
4,682,370 A	7/1987	Matthews
4,697,209 A	9/1987	Kiewit
4,739,398 A	4/1988	Thomas et al.
4,776,017 A	10/1988	Fujimoto
4,805,020 A	2/1989	Greenberg

(Continued)

## OTHER PUBLICATIONS

Peter N. Yianilos, Excluded Middle Vantage Point Forest for Nearest Neighbor Search, Aug. 1, 1999, pp. 1-12.\*

(Continued)

*Primary Examiner* — Brian Pendleton

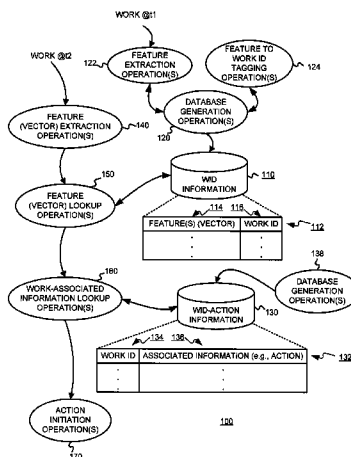
Assistant Examiner — Cai Chen

(74) *Attorney, Agent, or Firm* — Amster, Rothstein & Ebenstein LLP

(57) **ABSTRACT**

A media work may be associated with an action by (a) extracting features from the media work, (b) determining an identification of the media work, based on the features extracted, using a sub-linear time search, such as an approximate nearest neighbor search for example, and (c) determining an action based on the identification of the media work determined. The media work may be an audio work. The features extracted from the work may include (A) a frequency decomposition of a signal of the audio work, (B) information samples of the audio work, (C) average intensities of sampled windows of the audio work, and/or (D) information from frequencies of the audio work.

**40 Claims, 10 Drawing Sheets**



## US 8,205,237 B2

Page 2

## U.S. PATENT DOCUMENTS

4,843,526 A	6/1989	Price, III	7,523,312 B2	4/2009	Kalker et al.
4,843,562 A	6/1989	Kenyon et al.	7,587,728 B2	9/2009	Wheeler et al.
4,918,730 A	4/1990	Schulze	7,647,604 B2	1/2010	Ramaswamy
5,210,820 A	5/1993	Kenyon	7,650,616 B2	1/2010	Lee
5,283,819 A	2/1994	Glick et al.	7,757,248 B2	7/2010	Harkness et al.
5,437,050 A	7/1995	Lamb et al.	2001/0001160 A1 *	5/2001	Shoff et al. .... 725/51
5,481,294 A	1/1996	Thomas et al.	2001/0003818 A1	6/2001	Pingel et al.
5,581,658 A	12/1996	O'Hagan et al.	2002/0023020 A1	2/2002	Kenyon et al.
5,594,934 A	1/1997	Lu et al.	2002/0032698 A1	3/2002	Cox
5,629,739 A	5/1997	Dougherty	2002/0120925 A1	8/2002	Logan
5,692,213 A	11/1997	Goldberg et al.	2002/0156760 A1	10/2002	Lawrence et al.
5,701,452 A	12/1997	Siefert	2003/0106017 A1	6/2003	Kanchirayappa et al.
5,701,542 A	12/1997	Sasayama	2003/0146940 A1	8/2003	Ellis
5,724,605 A	3/1998	Wissner	2004/0199387 A1 *	10/2004	Wang et al. .... 704/243
5,745,900 A	4/1998	Burrows	2005/0160363 A1	7/2005	Bhagal et al.
5,798,785 A	8/1998	Hendricks et al.	2006/0101069 A1	5/2006	Bell et al.
5,850,490 A	12/1998	Johnson	2006/0206462 A1	9/2006	Barber
5,918,223 A	6/1999	Blum et al.	2007/0041667 A1	2/2007	Cox
5,953,415 A	9/1999	Nielsen	2007/0083510 A1	4/2007	McArdle
6,006,256 A	12/1999	Zdepski et al.	2007/0118375 A1	5/2007	Kenyon et al.
6,011,758 A	1/2000	Dockes et al.	2008/0091684 A1	4/2008	Ellis et al.
6,026,439 A	2/2000	Chowdhury et al.	2008/0250241 A1	10/2008	Ginter et al.
6,044,402 A	3/2000	Jacobson et al.			
6,052,693 A	4/2000	Smith et al.			
6,061,056 A *	5/2000	Menard et al. .... 715/704			
6,088,455 A	7/2000	Logan et al.			
6,088,707 A	7/2000	Bates et al.			
6,118,450 A	9/2000	Proehl et al.			
6,119,124 A	9/2000	Broder et al.			
6,169,986 B1	1/2001	Bowman et al.			
6,173,406 B1	1/2001	Wang et al.			
6,240,409 B1	5/2001	Aiken			
6,243,725 B1	6/2001	Hempleman et al.			
6,247,133 B1	6/2001	Palage et al.			
6,253,193 B1	6/2001	Ginter et al.			
6,263,348 B1	7/2001	Kathrow et al.			
6,330,593 B1	12/2001	Roberts et al.			
6,345,256 B1	2/2002	Milsted et al.			
6,349,296 B1	2/2002	Broder			
6,360,215 B1	3/2002	Judd et al.			
6,363,377 B1	3/2002	Kravets et al.			
6,374,225 B1	4/2002	Hejna, Jr.			
6,381,601 B1	4/2002	Fujiwara et al.			
6,385,596 B1	5/2002	Wiser et al.			
6,408,128 B1	6/2002	Abecassis			
6,418,421 B1	7/2002	Hurtado et al.			
6,446,068 B1	9/2002	Kortge			
6,449,226 B1	9/2002	Kumagai			
6,452,874 B1	9/2002	Otsuka et al.			
6,477,704 B1	11/2002	Cremia			
6,496,802 B1	12/2002	Van Zoest et al.			
6,505,160 B1	1/2003	Levy			
6,550,001 B1	4/2003	Corwin et al.			
6,550,011 B1	4/2003	Sims, III			
6,577,746 B1	6/2003	Evans et al.			
6,591,245 B1	7/2003	Klug			
6,598,228 B2	7/2003	Hejna, Jr.			
6,609,105 B2	8/2003	Van Zoest et al.			
6,654,757 B1	11/2003	Stern			
6,665,661 B1	12/2003	Crow et al.			
6,675,174 B1	1/2004	Bolle et al.			
6,834,308 B1 *	12/2004	Ikezoye et al. .... 709/231			
6,873,982 B1	3/2005	Bates et al.			
6,931,451 B1	8/2005	Logan et al.			
6,941,275 B1	9/2005	Swierczek			
6,978,419 B1	12/2005	Kantrowitz			
6,978,461 B2	12/2005	Shapiro et al.			
6,990,453 B2	1/2006	Wang et al.			
7,013,301 B2	3/2006	Holm et al.			
7,058,223 B2	6/2006	Cox			
7,106,904 B2	9/2006	Shima			
7,155,449 B2	12/2006	Pingel et al.			
7,158,929 B2	1/2007	Wouters et al.			
7,168,083 B2	1/2007	Kalker et al.			
7,302,574 B2	11/2007	Conwell et al.			
7,366,718 B1	4/2008	Pugh et al.			
7,421,723 B2	9/2008	Harkness et al.			
7,477,739 B2	1/2009	Haitsma et al.			

## OTHER PUBLICATIONS

Peter N. Yianilos, Excluded Middle Vantage Point Forest for Nearest Neighbor Search, Aug. 1, 1999, pp. 1-12.\*

P.N. Yianilos, "Locally Lifting the Curse of Dimensionality for Nearest Neighbor Search" *SODA 2000*, pp. 361-370.

Baum, L., et al., "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains", *The Annals of Mathematical Statistics*, vol. 41, No. 1, pp. 164-171 (1970).

Dempster, A. P., et al., "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, Issue 1, pp. 1-38 (1977).

Reynolds, D., et al., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Transactions on Speech and Audio Processing*, vol. 3, No. 1, pp. 72-83 (Jan. 1995).

Nievergelt, J. et al., "The Grid File: An Adaptable, Symmetric Multikey File Structure," *ACM Transactions on Database Systems*, vol. 9, No. 1, pp. 38-71 (Mar. 1984).

Heintze, N., "Scalable Document Fingerprinting," *Proc. USENIX Workshop on Electronic Commerce* (1996).

Wold, E. et al., "Content-Based Classification, Search, and Retrieval of Audio," *IEEE Multimedia*, vol. 3, Issue 3, pp. 27-63 (1996).

Bhanu, B., et al., "Learning Feature Relevance and Similarity Metrics in Image Databases", *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, pp. 14-19 (1998).

Del Bimbo, A., et al., "Using Weighted Spatial Relationships in Retrieval by Visual Contents", *Image Description and Retrieval*, pp. 161-192 (1998).

Indyk, P., and Motwani, R., "Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality," *Proceeding of the Thirtieth Annual ACM Symposium on Theory of Computing*, pp. 604-613 (1998).

La Cascia, M., et al., "Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web", *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, pp. 24-29 (1998).

Yoshitaka, A., et al., "A Survey on Content-Based Retrieval for Multimedia Databases", *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, No. 1, pp. 81-93 (Jan./Feb. 1999).

Lawrence, S., et al., "Digital Libraries and Autonomous Citation Indexing," *IEEE Computer*, pp. 67-71 (Jun. 1999).

Kimura, A. et al., "Very Quick Audio Searching: Introducing Global Pruning to the Time-Series Active Search," *IEEE Conf on Acoustics, Speech and Signal Processing, (ICASSP '01)*, vol. 3, pp. 1429-1432 (2001).

Chavez, E., et al., "Searching in Metric Spaces", *ACM Computing Surveys*, vol. 33, No. 3, pp. 273-321 (Sep. 2001).

Haitsma, J., et al., "Robust Audio Hashing for Content Identification, Int' Workshop on Content Based Multimedia Indexing, Brescia, Italy (Sep. 19-21, 2001).

## US 8,205,237 B2

Page 3

- Haitsma, J., and Walker, T., "A Highly Robust Audio Fingerprinting System," *Journal of New Music Research*, 1744-5027, vol. 32, Issue 2, pp. 211-221 (2003).
- Schleimer, Saul, et al., "Winnowing: Local Algorithms for Document Fingerprinting ACM SIGMOD" (Jun. 9-12, 2003).
- "Searching Near-Replicas of Images via Clustering" Edward Chang, Chen Li, James Wang, Peter Mork, Gio Wiederhold Proc. SPIE Symposium of Voice, Video, and Data Communications, 1999.
- "RIME: A Replicated Image Detector for the World-Wide Web" Edward Y. Chang, James Ze Wang, Chen Li, and Gio Wiederhold, SPIE 1998.
- "Safeguarding and charging for information on the internet," H. Garcia-Molina, S. Ketchpel, and N. Shivakumar, Proceedings of ICDE, 1998.
- "Detection mechanisms for digital documents," S. Brin and H. Garcia-Molina, Proceedings of ACM SIG-MOD, May 1995.
- "The x-tree: An index structure for high-dimensional data," S. Berchtold, Proceedings of the 22nd VLDB, Aug. 1996.
- "The sr-tree: An index structure for high-dimensional nearest neighbor queries," N. Katayama and S. Satoh, Proceedings of ACM SIGMOD, May 1997.
- "The k-d-b-tree: A search structure for large multidimensional dynamic indexes," J. T. Robinson, Proceedings of ACM SIGMOD, Apr. 1981.
- "Query by image and video content: The QBIC system," M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, and et al, IEEE Computer 28(9), pp. 23{32, 1995.
- "Visual information retrieval," A. Gupta and R. Jain, Communications of the ACM 40(5), pp. 69-79, 1997.
- "Visualseek: A fully automated content-based image query system," J. R. Smith and S.-F. Chang, ACM Multimedia Conference, 1996.
- "Similarity indexing: Algorithms and performance," D. A. White and R. Jain, Proc. SPIE vol. 2670, San Diego, 1996.
- "The r\*-tree: an efficient and robust access method, for points and rectangles," N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, Proceedings of ACM Sigmod, May 1990.
- "R-trees: a dynamic index structure for spatial searching," A. Guttmann, Proceedings of ACM Sigmod, Jun. 1984.
- "Similarity indexing with the ss-tree," D. A. White and R. Jain, Proceedings of the 12th ICDE, Feb. 1996.
- "The tv-tree: an index structure for high-dimensional data," K.-L. Lin, H. V. Jagadish, and C. Faloutsos, VLDB Journal 3 (4), 1994.
- "M-tree: An efficient access method for similarity search in metric spaces," P. Ciaccia, M. Patella, and P. Zezula, Proceedings of the 23rd VLDB, Aug. 1997.
- "Nearest neighbor queries," N. Roussopoulos, S. Kelley, and F. Vincent, Proceedings of ACM Sigmod, May 1995.
- "An extensible hashing index for high-dimensional similarity search," C. Li, E. Chang, and J. Z. Wang, Stanford Technical Report, Aug. 1998.
- "Two algorithms for nearest-neighbor search in high dimensions" J. M. Kleinberg, Proc 29th STOC, 1997.
- "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996.
- "Adaptive Color Image Embeddings for Database Navigation" Yossi Rubner, Carlo Tomasi and Leonidas J. Guibas, Proceedings of the 1998 IEEE Asian Conference on Computer Vision.
- A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces R. Weber, H-J Schek, S. Blott Proc., 24th VLDB Conf. 1998.
- Bouktache, D., "A fast algorithm for the nearest neighbor classifier", IEEE Transactions on Pattern Analysis and Machine Intelligence, Mar. 1997, pp. 277-282.
- Nene et al., "A simple algorithm for nearest neighbor search in high dimensions", IEEE Transactions on Pattern Analysis and Machine Intelligence; Sep. 1997, pp. 989-1003.
- Arya et al. "Approximate nearest neighbor queries in fixed dimensions", Proceedings of the 4th annual ACM-SIAM Symposium on Discrete algorithms, 1993; pp. 271-280.
- K. Fukunaga and P. M. Narendra. A branch and bound algorithm for computing k-nearest neighbors. IEEE Trans. Comput., C-24:750{753, Jul. 1975.
- C.D. Feustel and L. G. Shapiro. The nearest neighbor problem in an abstract metric space. Pattern Recognition Letters, pp. 125{128, Dec. 1982.
- Dennis Shasha and Tsong-Li Wang. New techniques for best-match retrieval. ACM Transactions on Information Systems, 8(2):140{158, Apr. 1990.
- J. Uhlmann. Satisfying general proximity/similarity queries with metric trees. Information Processing Letters, 40 (4):175{9, Nov. 1991.
- Sergey Brin, "Near Neighbor Search in Large Metric Spaces", Proceedings of the 21st VLDB Conference, Zurich, Switzerland, Sep. 1995.
- D. P. Huttenlocher, et al. Comparing images using the hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(3):850{63, Sep. 1993.
- Seidl et al. "Optimal multi-step k-nearest neighbor search", Proceedings of ACM SIGMOD international conference on Managemet of data, 1998, pp. 154-165.
- W.A. Burkhard and R.M. Keller. Some Approaches to Best-Match File Searching. Communications of the ACM. vol. 16, No. 4, Apr. 1973.
- Kushilevitz et al. "Efficient search for approximate nearest neighbor in high dimensional spaces", Proceedings of the 30th annual ACM Symposium on Theory of computing, 1998, pp. 614-623. \* ;annual ACM Symposium on Theory of computing, 1998, pp. 614-623.
- Yianilos, P. "Data structures and algorithms for nearest neighbor search in general metric spaces", Proceedings of the ACM-SIAM Symposium on Discrete algorithms, 1993, pp. 311-321.
- Ardizzone, Edoardo et al., "Motion and Color-Based Video Indexing and Retrieval," Universita di palermo, Dipartimento di Ingegneria Elettrica, pp. 135-139, Viale delle Scienze, Palermo, Italy, IEEE 1996.
- Deng, Yining et al., "Content-based Search of Video Using Color, Texture, and Motion," Dept. of Electrical and Computer Engineering, University of California, Santa Barbara, CA, pp. 534-537, IEEE 1997.
- Fang, Min et al., "Computing Iceberg Queries Efficiently," Dept. of Computer Science, Stanford, CA, Paper No. 234, pp. 1-25.
- Flickner, Myron et al., "Query by Image and Video Content: The QBIC System," IBM Almaden Research Center, Sep. 1995, pp. 23-32, IEEE 1995.
- Gargi, U et al., "Performance Characterization and Comparison of Video Indexing Algorithms," Dept. of Computer Science and Engineering, The Pennsylvania State University, University Park, PA.
- Gionis, Aristides et al., "Similarity Search in High Demensions via Hashing," Dept. of Computer Science, Stanford University, Stanford, CA, pp. 518-529, Proceeding the 25th VLDB Conference, Edinburgh, Scotland, 1999.
- Indyk, Piotr et al., "Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality" (preliminary version) Dept. of Computer Science, Stanford University, Stanford, CA, pp. 1-13 & i-vii, Jul. 21, 1999.
- Iyengar, Giridharan et al., "Models for automatic classification of video sequences," MIT Media Laboratory, Cambridge, MA.
- Jain, Anil K., et al., "Image Retrieval using Color and Shape," Dept. of Computer Science, Michigan State University, Eas Lansing, MI, pp. 1-24, May 15, 1995.
- Ogle, Virginia E., et al., "Chabot: Retrieval from a Relational Database of Images," University of California at Berkeley, Computer pp. 40-48, IEEE 1995.
- Pentland, A. et al., "Photobook: Content-Based Manipulation of Image Databases," Perceptual Computing Section, The Media Laboratory, Massachusetts Institute of Tech., International Jorunal of Computer Vision 18(3), pp. 233-254 (1996), 1996 Kluwer Academic Publishers. Manuf. in The Netherlands.
- Shivakumar, Narayanan et al., "SCAM: A Copy Detection Mechanism for Digital Documents," Dept. of Computer Science, Stanford University, Stanford, CA, pp. 1-13.

**US 8,205,237 B2**

Page 4

---

Shivakumar, Narayanan et al., "Building a Scalable and Accurate Copy Detection Mechanism," Dept. of Computer Science, Stanford University, Stanford, CA.

Srihari, Rohini K., "Automatic Indexing and Content-Based Retrieval of Captioned Images," State University of New York, Buffalo, Theme Feature, pp. 49-56, Sep. 1995, IEEE 1995.

Swain, Michael and Ballard, Dana H., "Color Indexing," International Journal of Computer Vision 7:1, p. 11-32 (1991), 1991 Kluwer Academic Publishers. Manuf. in The Netherlands.

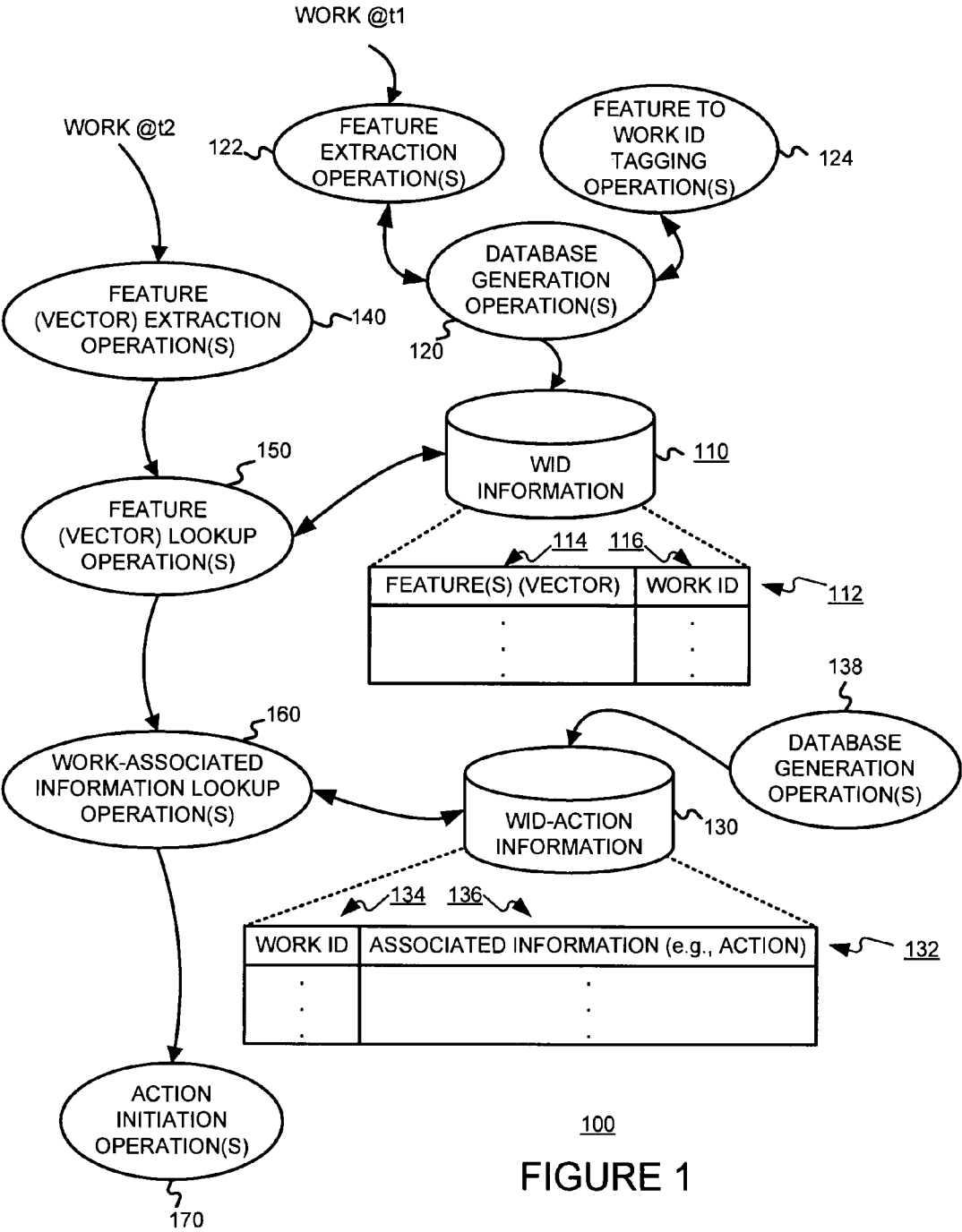
Wactlar, Howard D. et al., "Intelligence Access to Digital Video: Informedia Project," Carnegie Mellon University, Digital Library

Initiative: Carnegie Mellon University, Computer, pp. 46-52, IEEE 1996.

Yeo, Boon-Lock et al., "Rapid Scene Analysis on Compressed Video," IEEE Transactions on Circuits and Systems for Video Technology, vol. 5, No. 6, pp. 533-544, Dec. 1995, Dept. of Electrical Engineering, Princeton University, Princeton, NJ, IEEE Log No. 9415901.

Indyk, Piotr et al., "Finding pirated video sequences on the Internet," Dept. of Computer Science, Stanford University, Palo Alto, CA, Paper No. 199.

\* cited by examiner



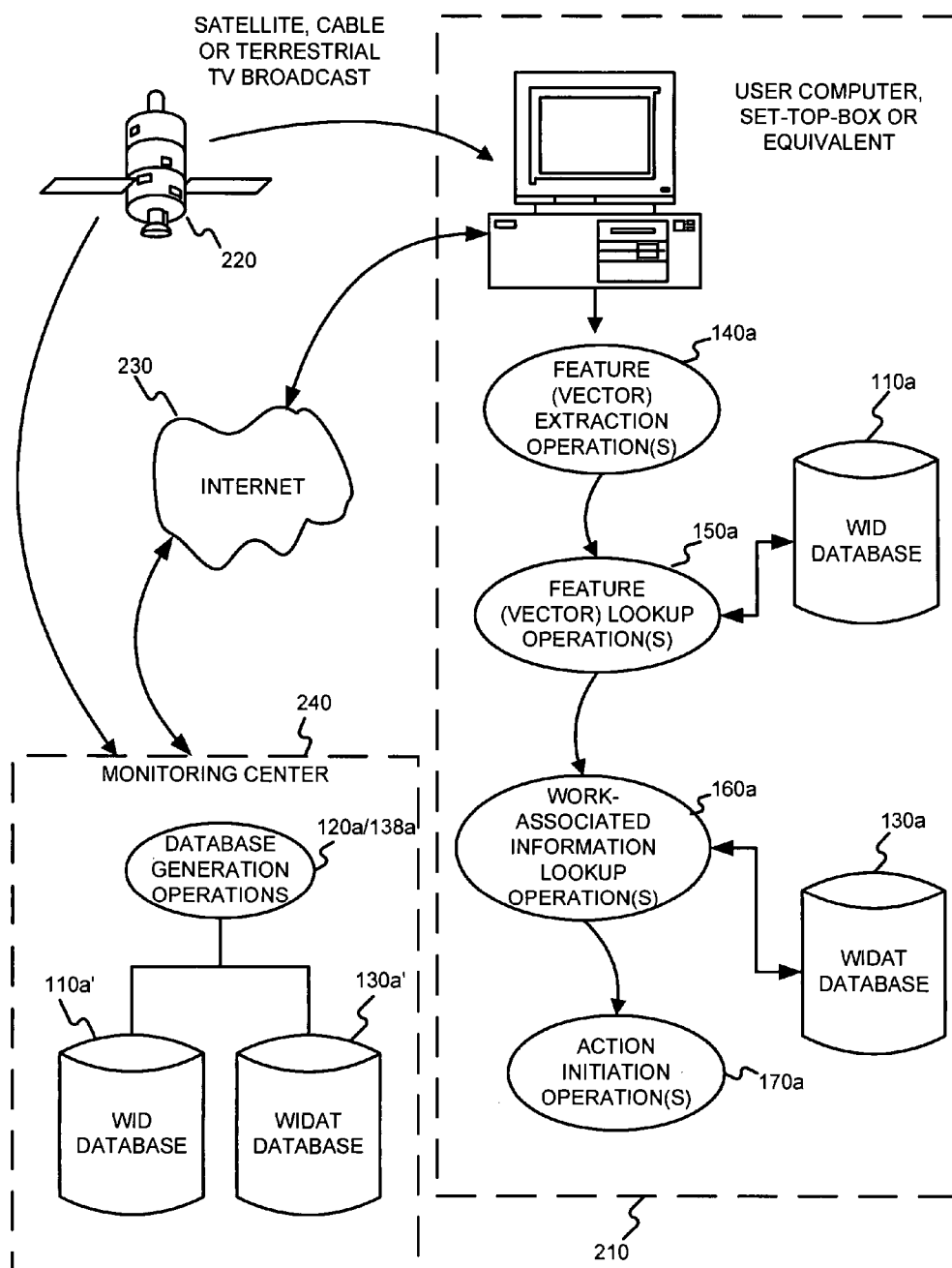


FIGURE 2

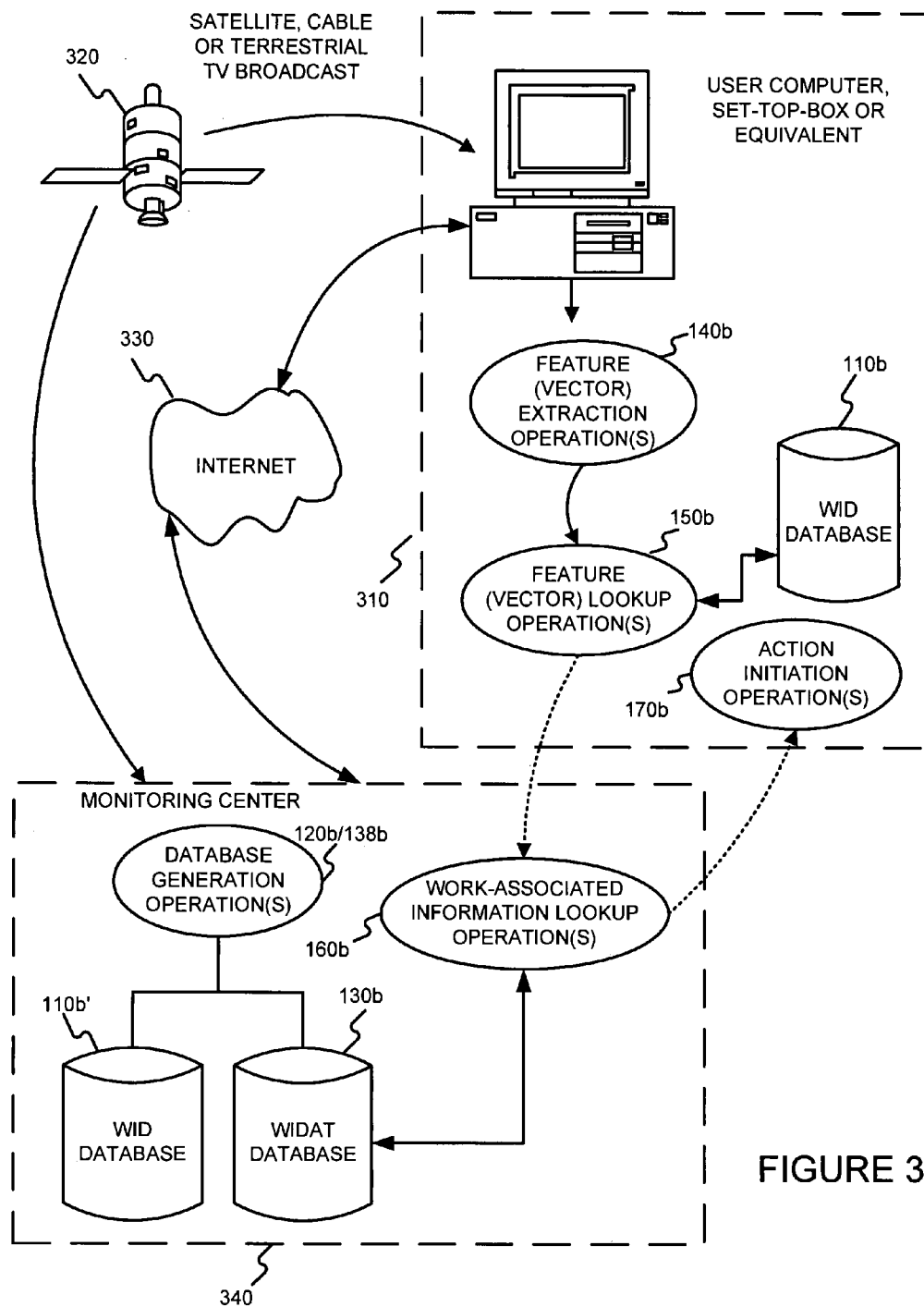


FIGURE 3



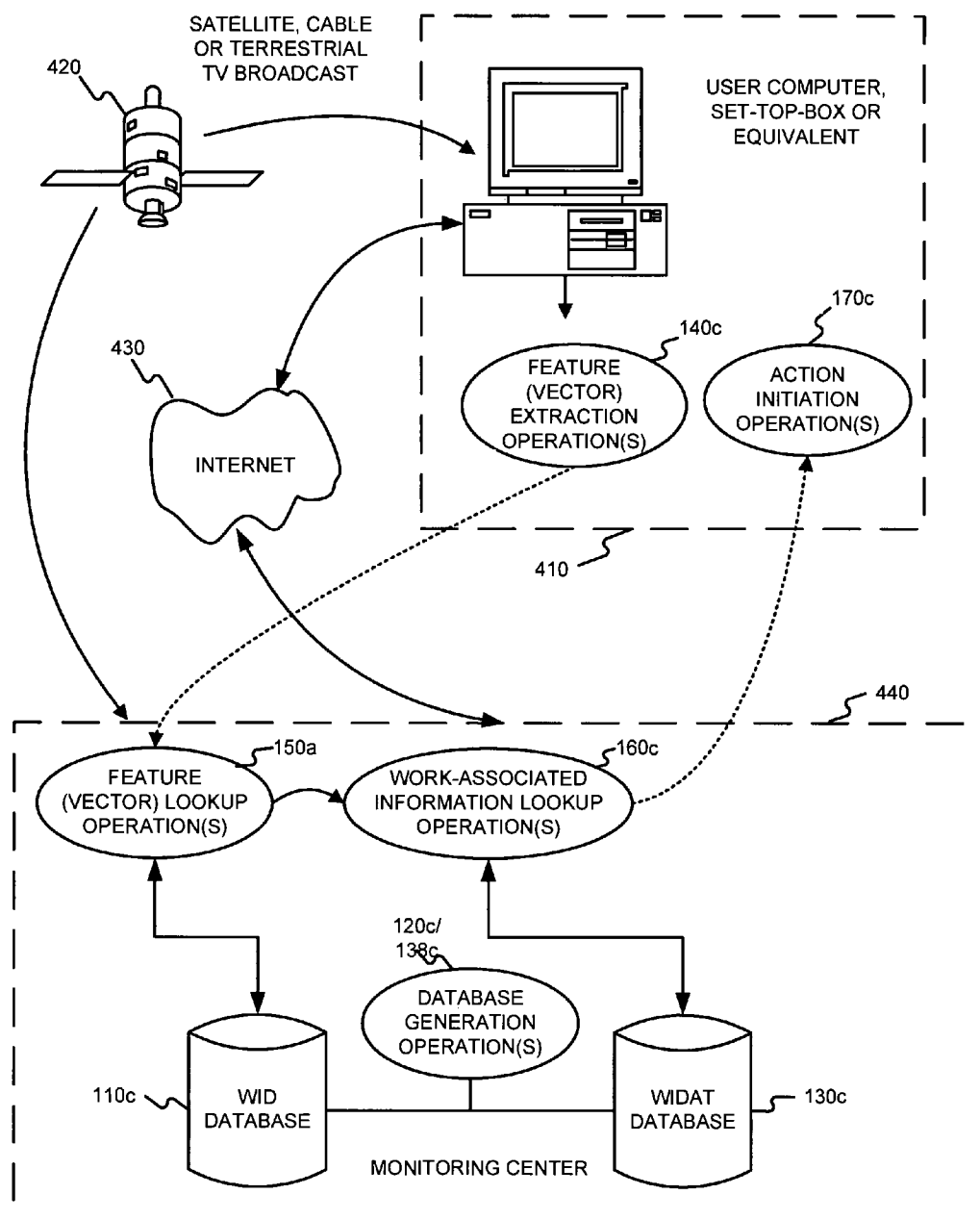


FIGURE 4

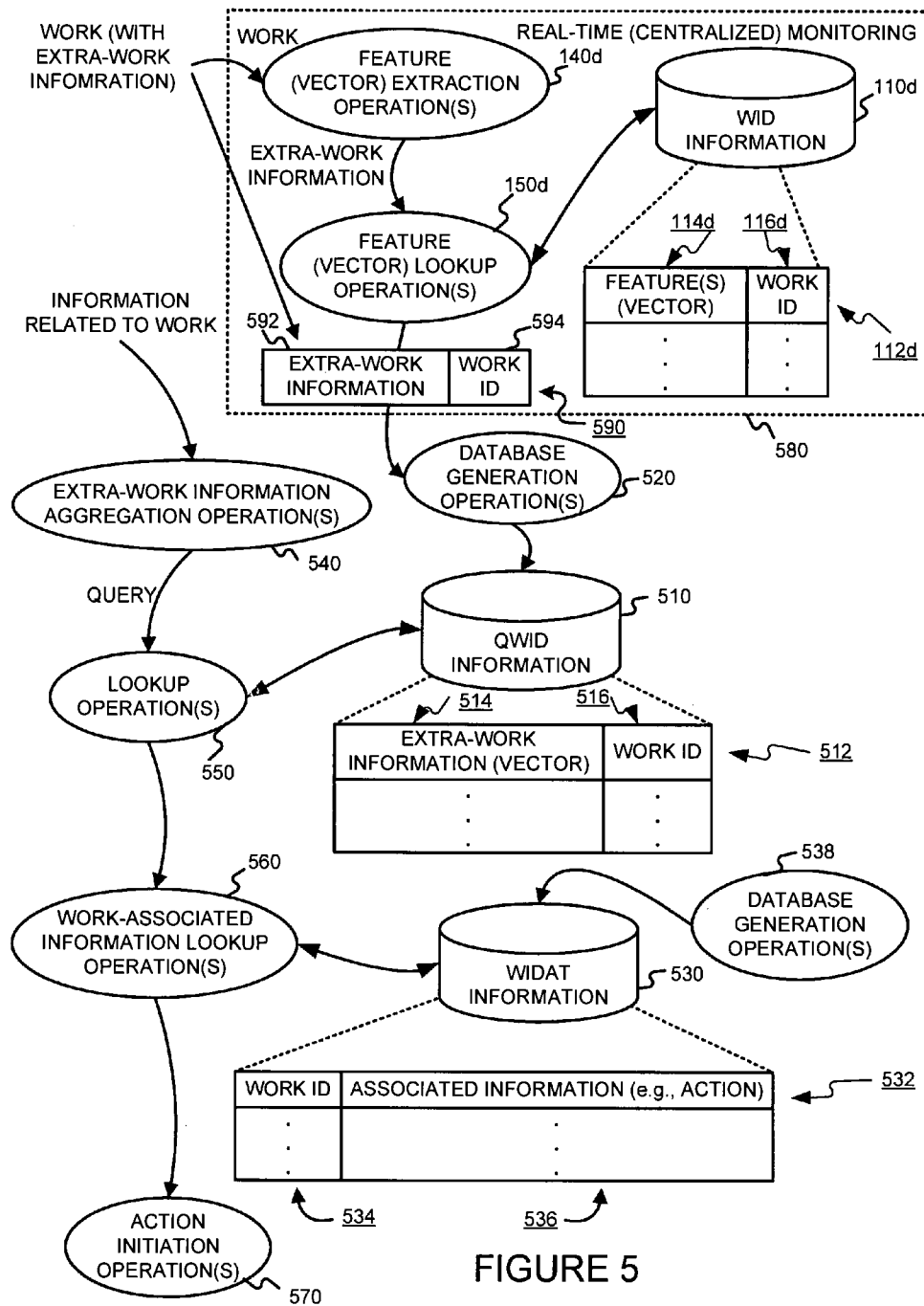


FIGURE 5

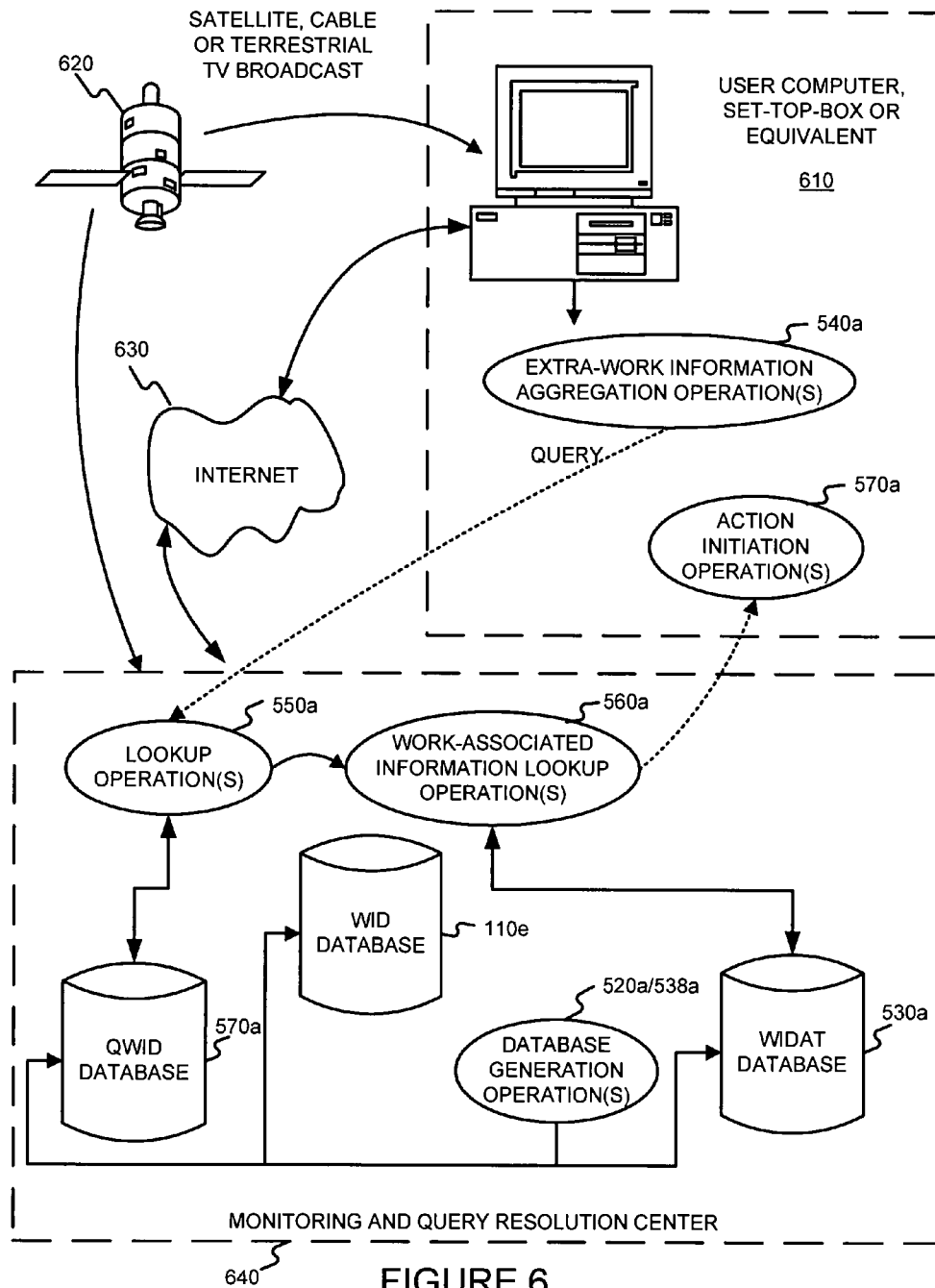


FIGURE 6

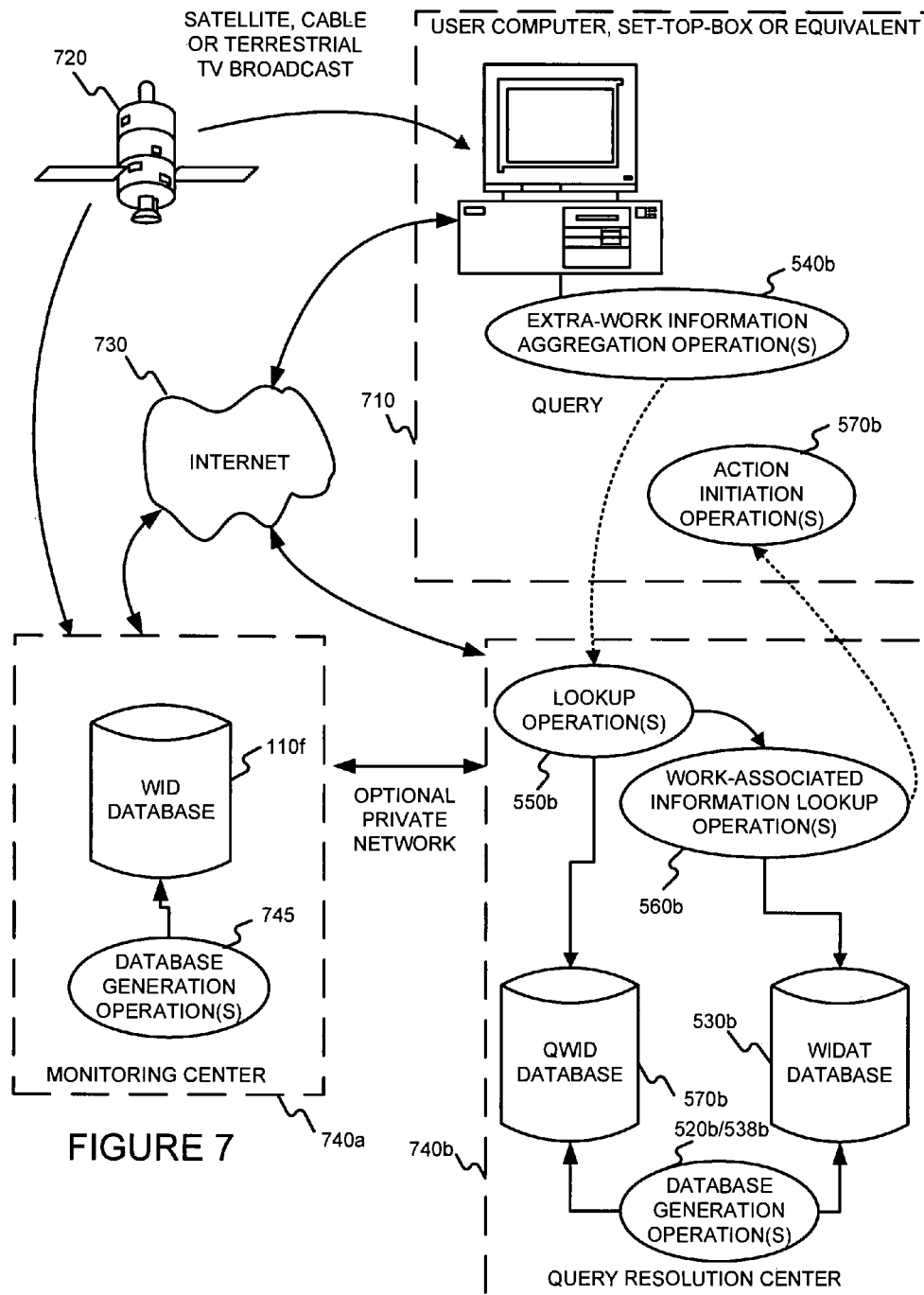


FIGURE 7

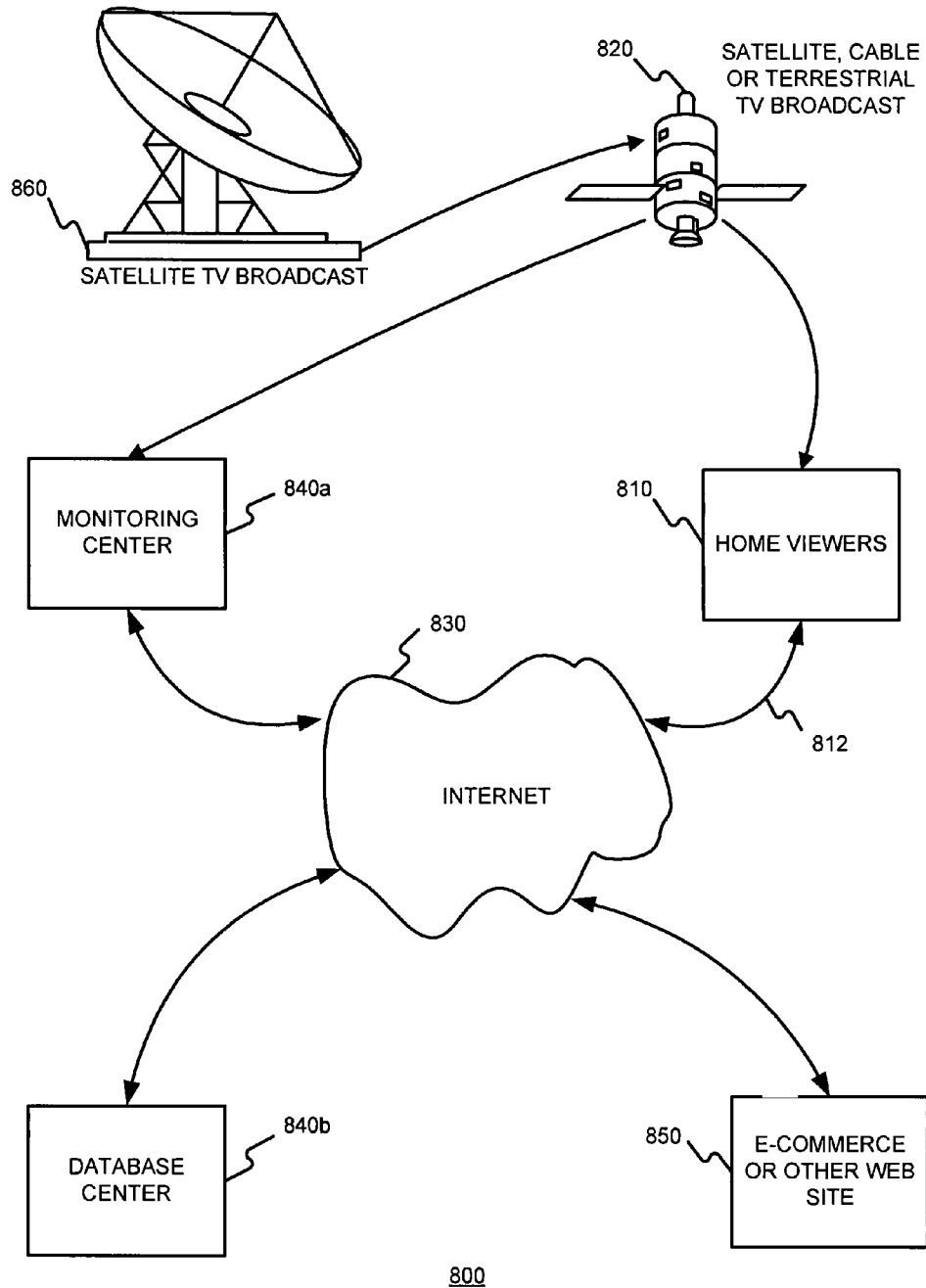


FIGURE 8

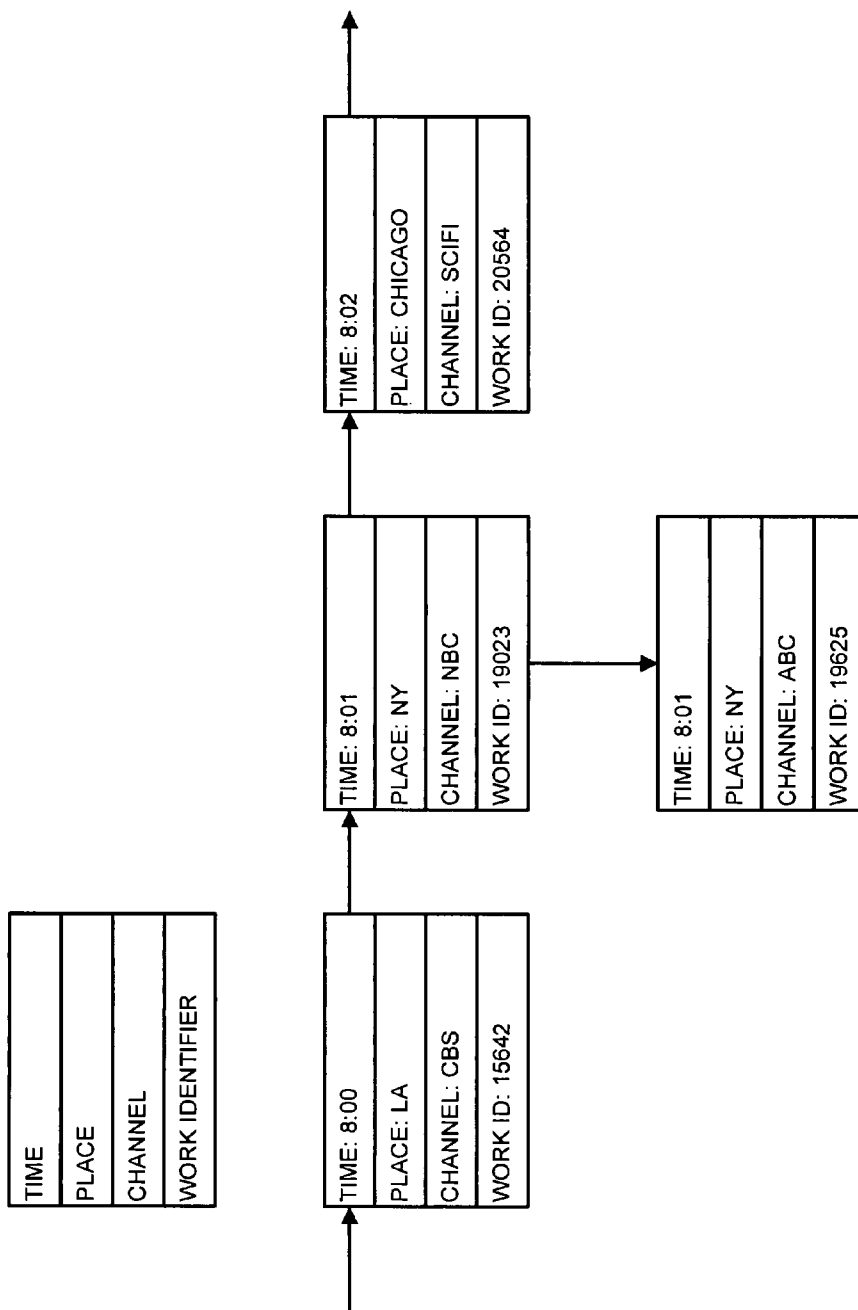
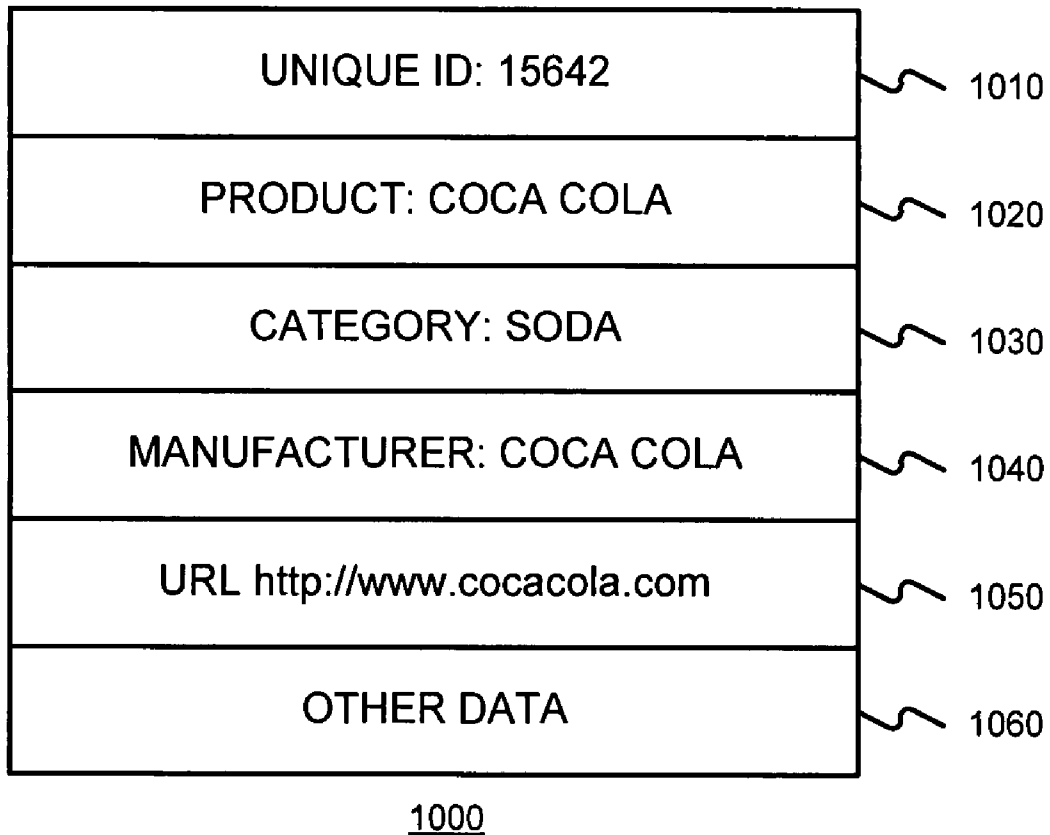


FIGURE 9



**FIGURE 10**

US 8,205,237 B2

1

**IDENTIFYING WORKS, USING A  
SUB-LINEAR TIME SEARCH, SUCH AS AN  
APPROXIMATE NEAREST NEIGHBOR  
SEARCH, FOR INITIATING A WORK-BASED  
ACTION, SUCH AS AN ACTION ON THE  
INTERNET**

§0. RELATED APPLICATIONS

The present application is a continuation of U.S. patent application Ser. No. 11/445,928 (incorporated herein by reference), titled "USING FEATURES EXTRACTED FROM AN AUDIO AND/OR VIDEO WORK TO OBTAIN INFORMATION ABOUT THE WORK," filed on Jun. 2, 2006, and listing Ingemar J. Cox as the inventor, which is a continuation-in-part of U.S. patent application Ser. No. 09/950,972 (incorporated herein by reference, issued as U.S. Pat. No. 7,058,223 on Jun. 6, 2006), titled "IDENTIFYING WORKS FOR INITIATING A WORK-BASED ACTION, SUCH AS AN ACTION ON THE INTERNET," filed on Sep. 13, 2001, now U.S. Pat. No. 7,058,223 and listing Ingemar J. Cox as the inventor, which application claims benefit to the filing date of provisional patent application Ser. No. 60/232,618 (incorporated herein by reference), titled "Identifying and linking television, audio, print and other media to the Internet", filed on Sep. 14, 2000 and listing Ingemar J. Cox as the inventor.

§1. BACKGROUND OF THE INVENTION

§1.1 Field of the Invention

The present invention concerns linking traditional media to new interactive media, such as that provided over the Internet for example. In particular, the present invention concerns identifying a work (e.g., content or an advertisement delivered via print media, or via a radio or television broadcast) without the need to modify the work.

§1.2 Related Art

§1.2.1 Opportunities Arising from Linking Works Delivered Via Some Traditional Media Channel or Conduit to a More Interactive System

The rapid adoption of the Internet and associated World Wide Web has recently spurred interest in linking works, delivered via traditional media channels or conduits, to a more interactive system, such as the Internet for example. Basically, such linking can be used to (a) promote commerce, such as e-commerce, and/or (b) enhance interest in the work itself by facilitating audience interaction or participation. Commerce opportunities include, for example, facilitating the placement of direct orders for products, providing product coupons, providing further information related to a product, product placement, etc.

In the context of e-commerce, viewers could request discount vouchers or coupons for viewed products that are redeemable at the point of purchase. E-commerce applications also extend beyond advertisements. It is now common for television shows to include product placements. For example, an actor might drink a Coke rather than a Pepsi brand of soda, actors and actresses might wear designer-labeled clothing such as Calvin Klein, etc. Viewers may wish to purchase similar clothing but may not necessarily be able to identify the designer or the particular style directly from the show. However, with an interactive capability, viewers would be able to discover this and other information by going to an associated Web site. The link to this Web site can be automatically enabled using the invention described herein.

In the context of facilitating audience interaction or participation, there is much interest in the convergence of tele-

2

vision and computers. Convergence encompasses a very wide range of capabilities. Although a significant effort is being directed to video-on-demand applications, in which there is a unique video stream for each user of the service, as well as to transmitting video signals over the Internet, there is also interest in enhancing the television viewing experience. To this end, there have been a number of experiments with interactive television in which viewers can participate in a live broadcast. There are a variety of ways in which viewers can participate. For example, during game shows, users can answer the questions and their scores can be tabulated. In recent reality-based programming such as the ABC television game show, "Big Brother", viewers can vote on contestants who must leave the show, and be eliminated from the competition.

§1.2.2 Embedding Work Identifying Code or Signals Within Works

Known techniques of linking works delivered via traditional media channels to a more interactive system typically require some type of code, used to identify the work, to be inserted into the work before it is delivered via such traditional media channels. Some examples of such inserted code include (i) signals inserted into the vertical blanking interval ("VBI") lines of a (e.g., NTSC) television signal, (ii) watermarks embedded into images, (iii) bar codes imposed on images, and (iv) tones embedded into music.

The common technical theme of these proposed implementations is the insertion of visible or invisible signals into the media that can be decoded by a computer. These signals can contain a variety of information. In its most direct form, the signal may directly encode the URL of the associated Web site. However, since the alphanumeric string has variable length and is not a particularly efficient coding, it is more common to encode a unique ID. The computer then accesses a database, which is usually proprietary, and matches the ID with the associated web address. This database can be considered a form of domain name server, similar to those already deployed for network addresses. However, in this case, the domain name server is proprietary and the addresses are unique ID's.

There are two principal advantages to encoding a proprietary identifier into content. First, as previously mentioned, it is a more efficient use of the available bandwidth and second, by directing all traffic to a single Web site that contains the database, a company can maintain control over the technology and gather useful statistics that may then be sold to advertisers and publishers.

As an example of inserting signals into the vertical blanking interval lines of a television signal, RespondTV of San Francisco, Calif. embeds identification information into the vertical blanking interval of the television signal. The VBI is part of the analog video broadcast that is not visible to television viewers. For digital television, it may be possible to encode the information in, for example, the motion picture experts group ("MPEG") header. In the USA, the vertical blanking interval is currently used to transmit close-captioning information as well as other information, while in the UK, the VBI is used to transmit teletext information. Although the close captioning information is guaranteed to be transmitted into the home in America, unfortunately, other information is not. This is because ownership of the vertical blanking interval is disputed by content owners, broadcasters and local television operators.

As an example of embedding watermarks into images, Digimarc of Tualatin, Oreg. embeds watermarks in print media. Invisible watermarks are newer than VBI insertion, and have the advantage of being independent of the method of



US 8,205,237 B2

3

broadcast. Thus, once the information is embedded, it should remain readable whether the video is transmitted in NTSC, PAL or SECAM analog formats or newer digital formats. It should be more reliable than using the vertical blanking interval in television applications. Unfortunately, however, watermarks still require modification of the broadcast signal which is problematic for a number of economic, logistical, legal (permission to alter the content is needed) and quality control (the content may be degraded by the addition of a watermark) reasons.

As an example of imposing bar codes on images, print advertisers are currently testing a technology that allows an advertisement to be shown to a camera, scanner or bar code reader that is connected to a personal computer ("PC"). The captured image is then analyzed to determine an associated Web site that the PC's browser then accesses. For example, GoCode of Draper, Utah embeds small two-dimensional bar codes for print advertisements. The latter signal is read by inexpensive barcode readers that can be connected to a PC. AirClic of Blue Bell, Pa. provides a combination of barcode and wireless communication to enable wireless shopping through print media. A so-called "CueCat" reads bar codes printed in conjunction with advertisements and articles in Forbes magazine. Similar capabilities are being tested for television and audio media.

Machine-readable bar codes are one example of a visible signal. The advantage of this technology is that it is very mature. However, the fact that the signal is visible is often considered a disadvantage since it may detract from the aesthetic of the work delivered via a traditional media channel or conduit.

As an example of embedding tones into music, Digital Convergence of Dallas, Tex. proposes to embed identification codes into audible music tones broadcast with television signals.

All the foregoing techniques of inserting code into a work can be categorized as active techniques in that they must alter the existing signal, whether it is music, print, television or other media, such that an identification code is also present. There are several disadvantages that active systems share. First, there are aesthetic or fidelity issues associated with bar codes, audible tones and watermarks. More importantly, all media must be processed, before it is delivered to the end user, to contain these active signals. Even if a system is enthusiastically adopted, the logistics involved with inserting bar codes or watermarks into, say every printed advertisement, are formidable.

Further, even if the rate of adoption is very rapid, it nevertheless remains true that during the early deployment of the system, most works will not be tagged. Thus, consumers that are early-adopters will find that most media is not identified. At best, this is frustrating. At worst, the naïve user may conclude that the system is not reliable or does not work at all. This erroneous conclusion might have a very adverse effect on the adoption rate.

Further, not only must there be modification to the production process, but modifications must also be made to the equipment in a user's home. Again, using the example of watermarking of print media, a PC must be fitted with a camera and watermark detection software must be installed. In the case of television, the detection of the identification signal is likely to occur at the set-top-box—this is the equipment provided by the local cable television or satellite broadcasting company. In many cases, this may require modifications to the hardware, which is likely to be prohibitively expensive. For example, the audible tone used by Digital Convergence to recognize television content, must be fed

4

directly into a sound card in a PC. This requires a physical connection between the television and the PC, which may be expensive or at least inconvenient, and a sound card may have to be purchased.

#### §1.2.3 Unmet Needs

In view of the foregoing disadvantages of inserting an identification code into a work, thereby altering the existing signal, there is a need for techniques of identifying a work without the need of inserting an identification code into a work. Such an identification code can then be used to invoke a work-related action, such as work-related commerce methods and/or to increase audience interest by facilitating audience interaction and/or participation.

## §2. SUMMARY OF THE INVENTION

Some embodiments consistent with the present invention provide a computer-implemented method, apparatus, or computer-executable programs for linking a media work to an action. Such embodiments might (a) extract features from the media work, (b) determine an identification of the media work based on the features extracted using a sub-linear time search, such as an approximate nearest neighbor search for example, and (c) determine an action based on the identification of the media work determined. In some embodiments consistent with the present invention, the media work is an audio signal. The audio signal might be obtained from a broadcast, or an audio file format. In other embodiments consistent with the present invention, the media work is a video signal. The video signal might be obtained from a broadcast, or a video file format.

In some of the embodiments pertaining to audio files, the audio file might be an mp3 file or some other digital representation of an audio signal. The information might include a song title, an album title, and/or a performer name.

In some of the embodiments pertaining to video files, the video file might be an MPEG file or some other digital representation of a video signal. The video file might be a video work, and the information might include a title of the video work, a director of the video work, and names of performers in the video work.

## §3. BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a process bubble diagram of operations that may be performed in accordance with one version of the present invention, in which intra-work information is used to identify the work.

FIG. 2 is a block diagram illustrating a first embodiment of the present invention, in which intra-work information is used to identify the work.

FIG. 3 is a block diagram illustrating a second embodiment of the present invention, in which intra-work information is used to identify the work.

FIG. 4 is a block diagram illustrating a third embodiment of the present invention, in which intra-work information is used to identify the work.

FIG. 5 is a process bubble diagram of operations that may be performed in accordance with another version of the present invention, in which extra-work information is used to identify the work.

FIG. 6 is a block diagram illustrating a fourth embodiment of the present invention, in which extra-work information is used to identify the work.

FIG. 7 is a block diagram illustrating a fifth embodiment of the present invention, in which extra-work information is used to identify the work.

## US 8,205,237 B2

5

FIG. 8 is a block diagram illustrating an environment in which the present invention may operate.

FIG. 9 is an exemplary data structure in which extra-work information is associated with a work identifier.

FIG. 10 is an exemplary data structure including work-related actions.

## §4. DETAILED DESCRIPTION

The present invention may involve novel methods, apparatus and data structures for identifying works without the need of embedding signals therein. Once identified, such information can be used to determine a work-related action. The following description is presented to enable one skilled in the art to make and use the invention, and is provided in the context of particular embodiments and methods. Various modifications to the disclosed embodiments and methods will be apparent to those skilled in the art, and the general principles set forth below may be applied to other embodiments, methods and applications. Thus, the present invention is not intended to be limited to the embodiments and methods shown and the inventors regard their invention as the following disclosed methods, apparatus, data structures and any other patentable subject matter to the extent that they are patentable.

## §4.1 FUNCTIONS

The present invention functions to identify a work without the need of inserting an identification code into a work. The present invention may do so by (i) extracting features from the work to define a feature vector, and (ii) comparing the feature vector to feature vectors associated with identified works. Alternatively, or in addition, the present invention may do so by (i) accepting extra-work information, such as the time of a query or of a rendering of the work, the geographic location at which the work is rendered, and the station that the audience member has selected, and (ii) use such extra-work information to lookup an identification of the work. In either case, an identification code may be used to identify the work.

The present invention may then function to use such an identification code to initiate a work-related action, such as for work-related commerce methods and/or to increase audience interest by facilitating audience interaction and/or participation.

## §4.2 EMBODIMENTS

As just introduced in §4.1 above, the present invention may use intra-work information and/or extra-work information to identify a work. Once identified, such identification can be used to initiate an action, such as an action related to commerce, or facilitating audience participation or interaction. Exemplary embodiments of the present invention, in which work is recognized or identified based on intra-work information, are described in §4.2.1. Then, exemplary embodiments of the present invention, in which work is recognized or identified based on extra-work information, are described in §4.2.2.

## §4.2.1 Embodiments in Which Work is Recognized Based on Intra-Work Information

## Such as a Feature Vector

Operations related to this embodiment are described in §4.2.1.1 below. Then, various architectures which may be used to effect such operations are described in §4.2.1.2.

6

## §4.2.1.1 Operations and Exemplary Methods and Techniques for Effecting Such Operations

FIG. 1 is a process bubble diagram of operations that may be performed in accordance with one version of the present invention, in which intra-work information is used to identify the work. As shown, a work-identification information storage 110 may include a number of items or records 112. Each item or record 112 may associate a feature vector of a work 114 with a, preferably unique, work identifier 116. The work-identification information storage 110 may be generated by a database generation operation(s) 120 which may, in turn, use a feature extraction operation(s) 122 to extract features from a work at a first time ( $WORK_{@t1}$ ), as well as a feature-to-work identification tagging operation(s) 124.

Further, work identifier-action information storage 130 may include a number of items or records 132. Each item or record 132 may associate a, preferably unique, work identifier 134 with associated information 136, such as an action for example. The work identifier-action information storage 130 may be generated by a database generation operation(s) 138 which may, for example, accept manual entries.

As can be appreciated from the foregoing, the work-information storage 110 records 112 and the work identification-action 130 records 132 can be combined into a single record. That is, there need not be two databases. A single database is also possible in which the work identifier, or a feature vector extracted from the work, serves as a key and the associated field contains work-related information, such as a URL for example.

The feature extraction operation(s) 140 can accept a work, such as that being rendered by a user, at a second time ( $WORK_{@t2}$ ), and extract features from that work. The extracted features may be used to define a so-called feature vector.

The extracted features, e.g., as a feature vector, can be used by a feature (vector) lookup operation(s) 150 to search for a matching feature vector 114. If a match, or a match within a predetermined threshold is determined, then the associated work identifier 116 is read.

The read work identifier can then be used by a work-associated information lookup operation(s) 160 to retrieve associated information, such as an action, 136 associated with the work identifier. Such information 136 can then be passed to action initiation operation(s) 170 which can perform some action based on the associated information 136.

## §4.2.1.1.1 Exemplary Techniques for Feature Extraction

When the user initiates a request, the specific television or radio broadcast or printed commercial, each of which is referred to as a work, is first passed to the feature extraction operation. The work may be an image, an audio file or some portion of an audio signal or may be one or more frames or fields of a video signal, or a multimedia signal. The purpose of the feature extraction operation is to derive a compact representation of the work that can subsequently be used for the purpose of recognition. In the case of images and video, this feature vector might be a pseudo-random sample of pixels from the frame or a low-resolution copy of the frame or the average intensities of  $n \times n$  blocks of pixels. It might also be a frequency-based decomposition of the signal, such as produced by the Fourier, wavelet and or discrete cosine transforms. It might involve principal component analysis. It might also be a combination of these. For television and audio signals, recognition might also rely on a temporal sequence of

feature vectors. The recognition literature contains many different representations. For block-based methods, blocks may be accessed at pseudo-random locations in each frame or might have a specific structure. For audio, common feature vectors are based on Fourier frequency decompositions, but other representations are possible. See, e.g., R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley-Interscience, New York, 1973). See also K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd Ed. (Academic Press, New York, 1990). (These references are incorporated herein by reference.)

As previously stated, one object of the vector extraction stage is to obtain a more concise representation of the frame. For example, each video frame is initially composed of 480×720 pixels which is equivalent to 345,600 pixels or 691,200 bytes. In comparison, an exemplary feature vector might only consist of 1 Kbyte of data.

A second purpose of the feature extraction process is to acquire a representation that is robust or invariant to possible noise or distortions that a signal might experience. For example, frames of a television broadcast may experience a small amount of jitter, i.e., horizontal and/or vertical translation, or may undergo lossy compression such as by MPEG-2. It is advantageous that these and other processes do not adversely affect the extracted vectors. For still images there has been considerable work on determining image properties that are invariant to affine and other geometric distortions. For example, the use of Radon and Fourier-Mellin transforms have been proposed for robustness against rotation, scale and translation, since these transforms are either invariant or bare a simple relation to the geometric distortions. See, e.g., C. Lin, M. Wu, Y. M. Lui, J. A. Bloom, M. L. Miller, I. J. Cox, "Rotation, Scale, and Translation Resilient Public Watermarking for Images," *IEEE Transactions on Image Processing* (2001). See also, U.S. Pat. Nos. 5,436,653, 5,504,518, 5,582,246, 5,612,729, and 5,621,454. (Each of these references is incorporated herein by reference.)

#### §4.2.1.1.2 Exemplary Techniques for Database Generation and Maintenance

A number of possibilities exist for generating and maintaining work identification (WID) and identification-action translation (WIDAT) databases. However, in all cases, works of interest are processed to extract a representative feature vector and this feature vector is assigned a unique identifier. This unique identifier is then entered into the work identification (WID) database 110 as well as into the WIDAT database 130 together with all the necessary associated data. This process is referred to as tagging. For example, in the case of an advertisement, the WIDAT database 130 might include the manufacturer (Ford), the product name (Taurus), a product category (automotive) and the URL associated with the Ford Taurus car together with the instruction to translate the query into the associated URL.

The determination of all works of interest and subsequent feature vector extraction and tagging depends on whether content owners are actively collaborating with the entity responsible for creating and maintaining the database. If there is no collaboration, then the database entity must collect all works of interest and process and tag them. While this is a significant effort, it is not overwhelming and is certainly commercially feasible. For example, competitive market research firms routinely tabulate all advertisements appearing in a very wide variety of print media. Newspapers and magazines can be scanned in and software algorithms can be applied to the images to identify likely advertisements. These

possible advertisements can then be compared with advertisements already in the WID database 110. If there is a match, nothing further need be done. If there is not a match, the image can be sent to a human to determine if the page does indeed contain an advertisement. If so, the operator can instruct the computer to extract the representative feature vector and assign it a unique identifier. Then, the operator can insert this information into the content identification database and as well as update the corresponding WIDAT database 130 with all the necessary associated data. This is continually performed as new magazines and papers include new advertisements to maintain the databases. This is a cost to the database entity. Television and radio broadcasts can also be monitored and, in fact, broadcast monitoring is currently performed by companies such as Nielsen Media research and Competitive Media Reporting. Television and radio broadcasts differ from print media in the real-time nature of the signals and the consequent desire for real-time recognition.

In many cases, advertisers, publishers and broadcasters may wish to collaborate with the database provider. In this case, feature extraction and annotation and/or extra-work information may be performed by the advertiser, advertisement agency, network and/or broadcaster and this information sent to the database provider to update the database. Clearly, this arrangement is preferable from the database provider's perspective. However, it is not essential.

#### §4.2.1.1.3. Exemplary Techniques for Matching Extracted Features with Database Entries

The extracted feature vector is then passed to a recognition (e.g., feature look-up) operation, during which, the vector is compared to entries of known vectors 114 in a content identification (WID) database 110. It is important to realize that the matching of extracted and known vectors is not equivalent to looking up a word in an electronic dictionary. Since the extracted vectors contain noise or distortions, binary search might not be possible. Instead, a statistical comparison is often made between an extracted vector and each stored vector. Common statistical measures include linear correlation and related measures such as correlation coefficient, but other methods can also be used including mutual information, Euclidean distance and Lp-norms. These measures provide a statistical measure of the confidence of the match. A threshold can be established, usually based on the required false positive and false negative rates, such that if the correlation output exceeds this threshold, then the extracted and known vectors are said to match. See, e.g., R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley-Interscience, New York, 1973). See also, U.S. Pat. No. 3,919,474 by W. D. Moon, R. J. Weiner, R. A. Hansen and R. N. Linde, entitled "Broadcast Signal Identification System". (Each of these references is incorporated herein by reference.)

If binary search was possible, then a database containing N vectors would require at most log(N) comparisons. Unfortunately, binary search is not possible when taking a noisy signal and trying to find the most similar reference signal. This problem is one of nearest neighbor search in a (high-dimensional) feature space. In previous work, it was not uncommon to perform a linear search of all N entries, perhaps halting the search when the first match is found. On average, this will require N/2 comparisons. If N is large, this search can be computationally very expensive.

Other forms of matching include those based on clustering, kd-trees, vantage point trees and excluded middle vantage point forests are possible and will be discussed in more detail later. See, e.g., P. N. Yianilos "Excluded Middle Vantage



Point Forests for nearest Neighbor Search”, *Presented at the Sixth DIMACS Implementation Challenge: Near Neighbor Searches workshop*, (Jan. 15, 1999). See also, P. N. Yianilos, “Locally lifting the curse of Dimensionality for nearest Neighbor Search” *SODA 2000*: 361-370. (Each of these references is incorporated herein by reference.) Thus, for example, a sub-linear search time can be achieved. Unlike the kd-tree method which finds the nearest neighbor with certainty, randomized constructions, like the one described in P. N. Yianilos, “Locally lifting the curse of Dimensionality for nearest Neighbor Search” *SODA 2000*: 361-370, that succeed with some specified probability may be used. One example of a sub-linear time search is an approximate nearest neighbor search. A nearest neighbor search always finds the closest point to the query. An approximate nearest neighbor search does not always find the closest point to the query. For example, it might do so with some probability, or it might provide any point within some small distance of the closest point.

If the extracted vector “matches” a known vector in the content identification database, then the work has been identified. Of course, there is the risk that the match is incorrect. This type of error is known as a false positive. The false positive rate can be reduced to any desired value, but at the expense of the false negative rate. A false negative occurs when the vector extracted from a work is not matched to the database even though the work is present in the database. There are several reasons why a work’s feature vector may fail to match a feature vector database entry. First, the recognition system may not be capable of 100% accuracy. Second, the extracted vector will often contain noise as a result of the transmission process. This noise may alter the values of a feature vector to the extent that a match is no longer possible.

Finally, there is the case where the observed work is not present in the database. In this case, the work can be sent to an operator for identification and insertion in the database.

#### §4.2.1.1.4 Exemplary Work Based Actions

Assuming that the work is correctly identified, then the identifier can be used to retrieve associated information from the second work identification-action translation (WIDAT) database **130** that contains information **136** associated with the particular work **134**. This information may simply be a corresponding URL address, in which case, the action can be considered to be a form of network address translation. However, in general, any information about the work could be stored therein, together with possible actions to be taken such as initiating an e-commerce transaction. After looking up the work identifier **134** in the WIDAT database **130**, an action is performed on behalf of the user, examples of which has been previously described.

In addition to using the system to allow audience members of a work to connect to associated sites on the Internet, a number of other uses are possible. First, the work identification database **130** allows competitive market research data to be collected (e.g., the action may include logging an event). For example, it is possible to determine how many commercials the Coca Cola Company in the Chicago market aired in the month of June. This information is valuable to competitors such as Pepsi. Thus, any company that developed a system as described above could also expect to generate revenue from competitive market research data that it gathers.

Advertisers often wish to ensure that they receive the advertising time that was purchased. To do so, they often hire commercial verification services to verify that the advertisement or commercial did indeed run at the expected time. To do

so, currently deployed systems by Nielsen and CMR embedded active signals in the advertisement prior to the broadcast. These signals are then detected by remote monitoring facilities that then report back to a central system which commercials were positively identified. See for example U.S. Pat. No. 5,629,739 by R. A. Dougherty entitled “Apparatus and method for injecting an ancillary signal into a low energy density portion of a color television frequency spectrum”, U.S. Pat. No. 4,025,851 by D. E. Haselwood and C. M. Solar entitled “Automatic monitor for programs broadcast”, U.S. Pat. No. 5,243,423 by J. P. DeJean, D. Lu and R. Weissman, entitled “Spread spectrum digital data transmission over TV video”, and U.S. Pat. No. 5,450,122 by L. D. Keene entitled “In-station television program encoding and monitoring system and method”. (Each of these patents is incorporated herein by reference.) Active systems are usually preferred for advertisement verification because the required recognition accuracy is difficult to achieve with passive systems. The passive monitoring system described herein supports commercial verification.

#### §4.2.1.2 Exemplary Architectures

Three alternative architectural embodiments in which the first technique may be employed are now described with reference to FIGS. 2, 3, and 4.

FIG. 2 is a block diagram illustrating a first embodiment of the present invention, in which intra-work information is used to identify the work and in which an audience member device **210**, such as a PC for example, receives and renders a work that is consumed by an audience member (user). At some point, the user may wish to perform a work-specific action such as traversing to an associated Web site. Upon initiation of this request, the computer **210** performs the operations **140a**, **150a**, **160a** and **170a**, such as those shown in FIG. 1. To reiterate, these operations include a feature extraction operation(s) **140a**, feature vector lookup or matching operation(s) **150a** in connection with items or records **112a** in a work-identification (WID) database **110a**. If a matching feature vector **114a** is found, the work-associated information lookup operation(s) **160a** can use the associated work identifier **116a** to accessing a work identification-action translation (WIDAT) database **130a** to retrieve associated information **136a**, possibly including determining what action should be performed.

As described above, the two databases might be integrated into a single database. However, conceptually, they are described here as separate.

An example illustrating operations that can occur in the first embodiment of FIG. 1, is now described. Consider a print application, in which say 10,000 advertisements are to be recognized that appear in national newspapers and magazines. If 1 Kbyte is required to store each feature vector then approximately 10 Mbytes of storage will be required for the work identification database **110a**. Such a size does not represent a serious problem, in either memory or disk space, to present personal computers.

An important issue then becomes recognition rate. While this may be problematic, all the images are two-dimensional—three-dimensional object recognition is not required. Of course, since a low cost camera captures the printed advertisement, there may be a number of geometric distortions that might be introduced together with noise. Nevertheless, the application is sufficiently constrained that adequate recognition rates should be achievable with current state-of-the-art computer vision algorithms. See, e.g., P. N. Yianilos “Excluded Middle Vantage Point Forests for nearest Neighbor

bor Search", Presented at the Sixth DIMACS Implementation Challenge: Near Neighbor Searches workshop, Jan. 15, 1999. See also, P. N. Yianilos "Locally lifting the curse of Dimensionality for nearest Neighbor Search" SODA 2000: 361-370. (Each of these references is incorporated herein by reference.) Thus, for example, a sub-linear search time can be achieved. Unlike the kd-tree method which finds the nearest neighbor with certainty, randomized constructions, like the one described in P. N. Yianilos, "Locally lifting the curse of Dimensionality for nearest Neighbor Search" SODA 2000: 361-370, that succeed with some specified probability may be used. One example of a sub-linear time search is an approximate nearest neighbor search. Estimates of the size of the WIDAT database **130a** depend on what associated information (recall fields **136**) is stored. If, for example, only a URL address is needed, about 20 characters can typically represent most URLs. Thus, the size of the WIDAT database **130a** would be less than 1 Mbyte.

The configuration just described with reference to FIG. 2 places all of the processing and data on each user's local machine **210**. A number of alternative embodiments, in which some or all of the storage and processing requirements are performed remotely, will be described shortly.

As new works are created and made publicly available, the databases residing on a user's local computer become obsolete. Just as the database provider **240** must continually update the databases in order to remain current, there is also a need to update local databases on devices at audience member premises. This update process can be performed over the Internet **230** in a manner very similar to how software is currently upgraded. It is not necessary to download an entirely new database although this is an option. Rather, only the changes need to be transmitted. During this update process, the user's computer **210** might also transmit information to a central monitoring center **240** informing it of which advertisements the computer user has queried. This type of information is valuable to both advertisers and publishers. Of course, care must be taken to ensure the privacy of individual users of the system. However, it is not necessary to know the identity of individual users for the system to work.

FIG. 3 is a block diagram illustrating a second embodiment of the present invention, in which intra-work information is used to identify the work. Although the WIDAT database can be quite small, as illustrated in the exemplary embodiment described above with respect to FIG. 2, there is still the problem of keeping this database current. While periodic updates of the local databases may be acceptable, they become unnecessary if the WIDAT database **130b** is at a remote location **340**. In this arrangement, illustrated in FIG. 3, after the local computer **310** identifies the work, it sends a query to the remote WIDAT database **130b**. The query may contain the work identifier. The remote site **340** may then return the associated information **136**. Although the remote WIDAT database **130b** needs to be updated by the database provider, this can be done very frequently without the need for communicating the updates to the local computers **310**.

The second embodiment is most similar to active systems in which an embedded signal is extracted and decoded and the identifier is used to interrogate a central database. Consequently it has many of the advantages of such systems, while avoiding the need to insert signals into all works. One such advantage, is that the database provider receives real-time information relating to users' access patterns.

The WIDAT database **130b** might physically reside at more than one location. In such a case, some requests will go to one site, and other requests will go to another. In this way, over-

loading of a single site by too many users can be avoided. Other load balancing techniques are also applicable.

FIG. 4 is a block diagram illustrating a third embodiment of the present invention, in which intra-work information is used to identify the work. Recall that the WIDAT database may be small relative to that work identification database (WID). As the size of the work recognition (WID) database increases, the foregoing embodiments may become impractical. Consider, for example, a music application in which it is desired to identify 100,000 song titles. If it is again assumed that a 1 Kbyte vector can uniquely represent each song, then on the order of 100 Mbytes is now needed. This size is comparable to large application programs such as Microsoft's Office 2000 suite. Although this still does not represent an inordinate amount of disk space, if this data needs to reside in memory at all times, then very few present machines will have adequate resources. Clearly, at some point, the proposed architectures scales to a point where requirements become impractical. In this case, a further modification to the architecture is possible.

Since the storage and searching of the work-identifier (WID) database require the most computation and storage, it may be more economical to perform these actions remotely. Thus, for example, if a user is playing an MP3 music file and wants to go to a corresponding website, the MP3 file is passed to an operation that determines one or more feature vectors. In the third embodiment, instead of performing the matching locally **410**, the one or more vectors are transmitted to a central site **440** at which is stored the WID and WIDAT databases **110c** and **130c** together with sufficiently powerful computers to resolve this request and those of other computer users. This configuration is illustrated in FIG. 4. Similarly, if a user is playing an MPEG or other video file and wants to initiate a work-related action, the video file is passed to an operation **140c** that extracts one or more feature vectors. The entire video file need not be processed. Rather, it may be sufficient to process only those frames in the temporal vicinity to the users request, i.e., to process the current frame and or some number of frames before and after the current frame, e.g. perhaps 100 frames in all. The extracted feature vector or feature vectors can then be transmitted to a central site **440** which can resolve the request.

After successfully matching the feature vector, the central site **440** can provide the user with information directly, or can direct the user to another Web site that contains the information the user wants. In cases where the recognition is ambiguous, the central site **440** might return information identifying one of several possible matches and allow the user to select the intended one.

The third embodiment is particularly attractive if the cost of extracting the feature vector is small. In this case, it becomes economical to have feature vector extraction **140c** in digital set-top-boxes and in video recorders **410**. The latter may be especially useful for the new generation of consumer digital video recorders such as those manufactured by TIVO and Replay TV. These devices already have access to the Internet via a phone line. Thus, when someone watching a recorded movie from television reacts to an advertisement, the video recorder would extract one or more feature vectors and transmit them to a central site **440**. This site **440** would determine if a match existed between the query vector and the database of pre-stored vectors **110c**. If a match is found, the central server **440** would transmit the associated information, which might include a Web site address or an 800 number for more traditional ordering, back to the audience user device **410**. Of course, a consumer device **410** such as a digital video recorder might also store personal information of the owner to facilitate online e-commerce. Such a device **410** could store

## US 8,205,237 B2

13

the owner's name, address, and credit card information and automatically transmit them to an on-line store to complete a purchase. Very little user interaction other than to authorize the purchase might be needed. This type of purchasing may be very convenient to consumers.

Another advantage of the third embodiment is that it obviates the need to update local databases while, at the same time, the centrally maintained databases can be kept current with very frequent updating.

#### §4.2.2 Embodiments in which Work is Recognized Based on Extra-Work Information

Operations related to this embodiment are described in §4.2.2.1 below. Then, various architectures which may be used to effect such operations are described in §4.2.2.2.

If the cost of extracting a feature vector is too large, then the cost of deploying any of the embodiments described in §4.2.1 above may be prohibitive. This is particularly likely in very cost sensitive consumer products, including set-top-boxes and next generation digital VCR's. Acknowledging this fact, a different technique, one that is particularly well suited for broadcasted media such as television and radio as well as to content published in magazines and newspapers, is now described. This technique relies on the fact that a work need not be identified by a feature vector extracted from the work (which is an example of "intra-work information"), but can also be identified by when and where it is published or broadcast (which are examples of "extra-work information")

An example serves to illustrate this point. Consider the scenario in which a viewer sees a television commercial and responds to it. The embodiments described in §4.2.1 above required the user device (e.g., a computer or set-top-box) 210/310/410 to extract a feature vector. Such an extracted vector was attempted to be matched to another feature vector(s), either locally, or at a remote site. In the embodiments using a remote site, if the central site is monitoring all television broadcasts, then the user's query does not need to include the feature vector. Instead, the query simply needs to identify the time, geographic location and the station that the viewer is watching. A central site can then determine which advertisement was airing at that moment and, once again, return the associated information. The same is true for radio broadcasts. Moreover, magazines and newspapers can also be handled in this manner. Here the query might include the name of the magazine, the month of publication and the page number.

##### §4.2.2.1 Operations and Exemplary Methods and Techniques for Effecting Such Operations

FIG. 5 is a process bubble diagram of operations that may be performed in accordance with another version of the present invention, in which extra-work information is used to identify the work. As shown, a query work-identification (QWID) information storage 510 may include a number of items or records 512. Each item or record 512 may associate extra-work information 514, related to the work, with a, preferably unique, work identifier 516. The query work-identification (QWID) information storage 510 may be generated by a database generation operation(s) 520.

Further, work identifier-action information (WIDAT) storage 530 may include a number of items or records 532. Each item or record 532 may associate a, preferably unique, work identifier 534 with associated information 536, such as an action for example. The work identifier-action (WIDAT)

14

information storage 530 may be generated by a database generation operation(s) 538 which may, for example, accept manual entries.

As can be appreciated from the foregoing, the query work-information (QWID) storage 510 records 512 and the work identification-action (WIDAT) storage 530 records 532 can be combined into a single record.

The extra-work information aggregation (e.g., query generation) operation(s) 540 can accept a information related to a work, such as the time of a user request or of a rendering of the work, the geographic location at which the work is rendered, and the station that the audience member has selected, and generate a query from such extra-work information.

The query including the extra-work information can be used by a lookup operation(s) 550 to search for a "matching" set of information 514. If a match, or a match within a predetermined threshold is determined, then the associated work identifier 516 is read.

The read work identifier can then be used by a work-associated information lookup operation(s) 560 to retrieve associated information, such as an action, 536 associated with the work identifier. Such information 536 can then be passed to action initiation operation(s) 570 which can perform some action based on the associated information 536.

If the extra-work information of a work is known (in advance), generating the query work identifier (QWID) information 510 is straight-forward. If this were always the case, an intra-work information-based recognition operation would not be needed. However, very often this is not the case. For example, local television broadcasts typically have discretion to insert local advertising, as well as national advertising. Thus, it often is not possible to know in advance when, on what station, and where a particular advertisement will play.

In such instances, a real-time (e.g., centralized) monitoring facility 580 may be used to (i) extract feature vectors from a work, (ii) determine a work identifier 116 from the extracted features, and (iii) communicate one or more messages 590 in which extra-work information (e.g., time, channel, geographic market) 592 is associated with a work identifier 594, to operation(s) 520 for generating query work identification (QWID) information 510.

##### §4.2.2.1.1 Exemplary Extra-Work Information

In the context of national broadcasts, geographic information may be needed to distinguish between, for example, the ABC television broadcast in Los Angeles and that in New York. While both locations broadcast ABC's programming, this programming airs at different times on the East and West coasts of America. More importantly, the local network affiliates that air ABC's shows have discretion to sell local advertising as well as a responsibility to broadcast the national commercials that ABC sells. In short, the works broadcast by ABC in Los Angeles can be different from that in other geographic locations. Geographic information is therefore useful to distinguish between the different television markets. In some circumstances, geographic information may not be necessary, especially in parts of the world with highly regulated and centralized broadcasting in which there are not regional differences.

##### §4.2.2.1.2 Exemplary Techniques for Generating Databases

FIG. 5 illustrates a third database 510 referred to as the query to work identification (QWID) database. This database



15

510 maps the query (e.g., in the form of time, location and channel information) into a unique ID that identifies the perceived work. The QWID 510 and WIDAT 530 databases might not be separate, but for clarity will be considered so. After retrieving the unique work identifier 512 from the QWID database 510, the identifier can be used to access the WIDAT database 530. This is discussed in more detail later.

As introduced above, although it appears that this architecture does not require a recognition facility, such a facility may be needed. The feature extraction operation(s) 140d, as well as the work identification operation(s) 150d and other databases 110d, may be moved to one or more remote sites 580.

Although TV Guide and other companies provide detailed information regarding what will be broadcast when, these scheduling guides do not have any information regarding what advertisements will air when. In many cases, this information is unknown until a day or so before the broadcast. Even then, the time slots that a broadcaster sells to an advertiser only provide a time range, e.g. 12 pm to 3 pm. Thus it is unlikely that all commercials and aired programming can be determined from TV schedules and other sources prior to transmission. Further, occasionally programming schedules are altered unexpectedly due to live broadcasts that overrun their time slots. This is common in sports events and awards shows. Another example of interrupts to scheduled programming occurs when a particularly important news event occurs.

During transmission, it may therefore be necessary for a central site 580 to determine what work is being broadcast and to update its and/or other's database 520 accordingly based on the work identified 594 and relevant extra-work information 592. There are a variety of ways that this can be accomplished.

First, it may be economically feasible to manually monitor all television stations that are of interest, and manually update the database with information regarding the work being monitored. In fact, Nielsen used such procedures in the early 1960's for the company to tabulate competitive market data. More than one person can be employed to watch the same channel in order to reduce the error rate. It should be noted that the recent ruling by the FCC that satellite broadcasters such as DirecTV, DishTV and EchoStar can carry local stations significantly reduces the cost of monitoring many geographic markets. Currently, DirecTV, for example, carries the four main local stations in each of the 35 largest markets. Thus, these  $4 \times 35 = 140$  channels can all be monitored from a single site 580. This site would be provided with satellite receivers to obtain the television channels.

Unfortunately, however, humans are error prone and the monitoring of many different stations from many different geographic locations can be expensive. In order to automate the recognition process, a central site 580 could employ a computer-based system to perform automatic recognition. Because the recognition is centralized, only one or a few sites are needed. This is in comparison with the first architecture we described in which a complete recognition system was required in every user's home or premise. This centralization makes it more economic to employ more expensive computers, perhaps even special purpose hardware, and more sophisticated software algorithms. When video frames or clips cannot be identified or are considered ambiguous, this video can be quickly passed to human viewers to identify. Further, it should be possible for the automated recognition system to use additional information such as television schedules, time of day, etc in order to improve its recognition rate.

#### §4.2.2.1.2 Exemplary Techniques for Generating Queries Based on Extra-Work Information

At the audience member (user) premises, all that is needed is for the device to send a query to a database-server with

16

information that includes extra-work information, such as geographic location, time and channel. Usually, this extra-work information would be transmitted in real-time, while the work (e.g., an advertisement) is being broadcast. However, this is not necessary. If the television does not have access to the Internet, and most TV's do not yet, then an audience member (user) may simply remember or record which channel he or she was viewing at what time. In fact, the user device could store this information for later retrieval by the user. At a convenient later time, the user might access the Internet using a home PC. At this time, he or she can query the database by entering this extra-work information (e.g., together with geographic information) into an application program or a web browser plug-in.

Another possibility is allowing an audience member (user), at the time he or she is consuming (e.g., viewing, reading, listening to, etc.) the work, to enter query information into a handheld personal digital assistant ("PDA") such as a Palm Pilot, so as not to forget it. This information can then be manually transferred to a device connected to a network, or the information can be transferred automatically using, for example, infrared communications or via a physical link such as a cradle. Recently, PDAs also have some wireless networking capabilities built in, and thus might support direct access to the information desired. Further, software is available that allows a Palm Pilot or other PDA to function as a TV remote control device. As such, the PDA already knows the time of day and channel being viewed. It also probably knows the location of the audience member, since most PDA users include their own name and address in the PDA's phonebook and identify it as their own. Thus, with one or a few clicks, an audience member PDA user could bookmark the television content he or she is viewing. If the PDA is networked, then the PDA can, itself, retrieve the associated information immediately. Otherwise, the PDA can transfer this bookmarked data to a networked device, which can then provide access to the central database.

#### §4.2.2.2 Exemplary Architectures

FIG. 6 is a block diagram illustrating a fourth embodiment of the present invention, in which extra-work information is used to identify the work. As shown, an extra-work information aggregation operation 540a may be effected on a device 610, such as a PC, at the audience member (user) premises. The various databases 510a, 530a, and 110e, as well as the database generation operation(s) 520a/538a, the lookup operation(s) 550a and the work-associated information lookup operation(s) 560a may be provided at one or more centralized monitoring and query resolution centers 640.

FIG. 7 is a block diagram illustrating a fifth embodiment of the present invention, in which extra-work information is used to identify the work. This fifth embodiment is similar to the fourth embodiment illustrated in FIG. 6 but here, the monitoring center 740a and query resolution center 740b are separate.

These embodiments have many advantages for television and radio broadcasters who desire to provide Internet links or other action. First, the audience member (user) equipment, whether it is a computer, set-top-box, television, radio, remote control, personal digital assistant (pda), cell phone or other device, does not need to perform any processing of the received signal. As such, there is almost no cost involved to equipment manufacturers.

These last embodiments have some similarity with services such as those provided by the companies Real Names of Redwood City, Calif., America Online ("AOL") and espe-

cially iTag from Xenote. The popular press has reported on the difficulties associated with assigning domain names. The simplest of these problems is that almost all the one-word names in the “.com” category have been used. Consequently, domain names can often be difficult to remember. To alleviate this problem, RealNames and AOL provide alternative, proprietary name spaces (AOL calls these keywords). For a fee, a company may register a name with these companies. Thus, rather than type the URL <http://www.bell-labs.com>, the simple keyword “bell” might be sufficient to access the same Web site. These capabilities are convenient to users. However, these systems are very different from the fourth and fifth embodiments described. First, and foremost, these systems are not designed to identify content. Rather, they are simply alternative network address translation systems based on easily remembered mnemonics which are sold to interested companies. As such, the user is still expected to type in an address, but this address is easier to remember than the equivalent URL. In contrast, while a user may manually enter the information describing the work, the preferred embodiment is for the computer, set-top-box or other device to automatically generate this information. Further, the mapping of keywords to network addresses is an arbitrary mapping maintained by AOL or Real Names. For example, the keyword “bell” might just as reasonably point to the Web site for Philadelphia’s Liberty Bell as to Lucent’s Bell Labs. In contrast, the query used in the fourth and fifth embodiments is designed to contain all the necessary data to identify the work, e.g. the time, place and television channel during which the work was broadcast. There is nothing arbitrary about this mapping. It should also be pointed out that the proposed system is dynamic—the same work, e.g. a commercial, potentially has an infinite number of addresses depending on when and where it is broadcast. If an advertisement airs 100,000 unique times, then there are 100,000 different queries that uniquely identify it. Moreover, the exemplary query includes naturally occurring information such as time, place, channel or page number. This is not the case for AOL or RealNames, which typically assigns one or more static keywords to the address of a Web site.

Xenote’s iTag system is designed to identify radio broadcasts and uses a query similar to that which may be used in the fourth and fifth embodiments, i.e. time and station information. However, the work identification information is not dynamically constructed but is instead based on detailed program scheduling that radio stations must provide it. As such, it suffers from potential errors in scheduling and requires the detailed cooperation of broadcasters. While the fourth and fifth embodiments might choose to use program scheduling information and other ancillary information to aid in the recognition process, they do not exclusively rely on this. The concept of resolving a site name by recognizing the content is absent from the above systems.

#### §4.2.3 Exemplary Apparatus for Audience Member (User) Premise Device

While personal computers may be the primary computational device at a user’s location, it is not essential to use a PC. This is especially true of the embodiments depicted in FIGS. 6 and 7, which do not require the content, e.g. video signal, to be processed. Instead, only a unique set of identification parameters such as time, location and channel are provided to identify the perceived Work. Many forms of devices can therefore take advantage of this configuration.

As previously noted, personal digital assistants (PDAs) can be used to record the identification information. This infor-

mation can then be transferred to a device with a network communication such as a PC. However, increasingly, PDAs will already have wireless network communication capabilities built-in, as with the Palm VII PDA. These devices will allow immediate communication with the query resolution center and all information will be downloaded to them or they can participate in facilitating an e-commerce transaction. Similarly, wireless telephones are increasingly offering web-enabled capabilities. Consequently, wireless phones could be programmed to act as a user interface.

New devices can also be envisaged, including a universal remote control for home entertainment systems with a LCD or other graphical display and a network connection. This connection may be wireless or the remote control might have a phone jack that allows it to be plugged directly into an existing phone line. As home networks begin to be deployed, such devices can be expected to communicate via an inexpensive interface to the home network and from there to access the Internet.

In many homes, it is not uncommon for a computer and television to be used simultaneously, perhaps in the same room. A person watching television could install a web browser plug-in or applet that would ask the user to identify his location and the station being watched. Then, periodically, every 20 seconds for example, the plug-in would update a list of web addresses that are relevant to the television programs being watched, including the commercials. The audience member would then simply click on the web address of interest to obtain further information. This has the advantage that the viewer does not have to guess the relevant address associated with a commercial and, in fact, can be directed to a more specialized address, such as [www.fordvehicles.com/ibv/taurus2kflash/flash.html](http://www.fordvehicles.com/ibv/taurus2kflash/flash.html), rather than the generic [www.ford.com](http://www.ford.com) site. Of course, this applet or plug-in could also provide the database entity with information regarding what is being accessed from where and at what time. This information, as noted earlier, is valuable to advertisers and broadcasters. For PC’s that have infra-red communication capabilities, it is straightforward to either control the home entertainment center from the PC or for the PC to decode the signals from a conventional remote control. Thus, as a user changes channels, the PC is able to automatically track the channel changes.

Recording devices such as analog VCR’s and newer digital recording devices can also be exploited in the embodiments depicted in FIGS. 6 and 7, especially if device also record the channel and time information for the recorded content. When a user initiates a query, the recorded time and channel, rather than the current time and channel, then form part of the identification information.

Digital set-top-boxes are also expected to exploit the capabilities described herein. In particular, such devices will have two-way communication capabilities and may even include cable modem capabilities of course, the two-way communication need not be over a television cable. For example, satellite set-top-boxes provide up-link communications via a telephone connection. Clearly, such devices provide a convenient location to enable the services described herein. Moreover, such services can be provided as part of the OpenCable and DOCSIS (data over cable service interface specification) initiatives.

#### §4.2.4 Information Retrieval Using Features Extracted from Audio and/or Video Works

Some embodiments consistent with the present invention provide a computer-implemented method, apparatus, or com-



puter-executable program for providing information about an audio file or (a video file) played on a device. Such embodiments might (a) extract features from the audio (or video) file, (b) communicate the features to a database, and (c) receive the information about the audio (or video) file from the database. In some embodiments consistent with the present invention, the act of extracting the features is performed by a microprocessor of the device, and/or a digital signal processor of the device. The received information might be rendered on an output (e.g., a monitor, a speaker, etc.) of the device. The received information might be stored (e.g., persistently) locally on the device. The information might be stored on a disk, or non-volatile memory.

In some of the embodiments pertaining to audio files, the audio file might be an mp3 file or some other digital representation of an audio signal. The information might include a song title, an album title, and/or a performer name.

In some of the embodiments pertaining to video files, the video file might be an MPEG file or some other digital representation of a video signal. The video file might be a video work, and the information might include a title of the video work, a director of the video work, and names of performers in the video work.

#### §4.3 OPERATIONAL EXAMPLES

An example illustrating operations of an exemplary embodiment of the present invention, that uses intra-work information to identify the work, is provided in §4.3.1. Then, an example illustrating operations of an exemplary embodiment of the present invention, that uses extra-work information to identify the work, is provided in §4.3.2.

##### §4.3.1 Operational Example where Intra-Work Information is Used to Identify the Work

A generic system for monitoring television commercials is now described. Obviously, the basic ideas extend beyond this specific application.

The process of recognition usually begins by recognizing the start of a commercial. This can be accomplished by looking for black video frames before and after a commercial. If a number of black frames are detected and subsequently a similar number are detected 30 seconds later, then there is a good chance that a commercial has aired and that others will follow. It is also well known that the average sound volume during commercials is higher than that for television shows and this too can be used as an indicator of a commercial. Other methods can also be used. The need to recognize the beginning of a commercial is not essential. However, without this stage, all television programming must be assumed to be commercials. As such, all video frames must be analyzed. The advantage of determining the presence of a commercial is that less video content must be processed. Since the percentage of advertising time is relatively small, this can lead to considerable savings. For example, commercials can be buffered and then subsequently processed while the television show is being broadcast. This reduces the real-time requirements of a system at the expense of buffering, which requires memory or disk space. Of course, for the applications envisioned herein, a real-time response to a user requires real-time processing.

Once it is determined that an advertisement is being broadcast, it is necessary to analyze the video frames. Typically, a compact representation of each frame is extracted. This vector might be a pseudo-random sample of pixels from the frame or a low-resolution copy of the frame or the average

intensities of  $n \times n$  blocks of pixels. It might also be a frequency-based decomposition of the signal, such as produced by the Fourier, Fourier-Mellin, wavelet and/or discrete cosine transforms. It might involve principal component analysis or any combination thereof. The recognition literature contains many different representations. For block-based methods, the  $n \times n$  blocks may be located at pseudo-random locations in each frame or might have a specific structure, e.g. a complete tiling of the frame. The feature vector might then be composed of the pixels in each block or some property of each block, e.g. the average intensity or a Fourier or other decomposition of the block. The object of the vector extraction stage is to obtain a more concise representation of the frame. Each frame is initially composed of  $480 \times 720$  pixels which is equivalent to 345,600 bytes, assuming one byte per pixel. In comparison, the feature vector might only consist of 1 Kbyte of data. For example, if each frame is completely tiled with  $16 \times 16$  blocks, then the number of blocks per frame is  $345,600/256=1350$ . If the average intensity of each block constitutes the feature vector, then the feature vector consists of 1350 bytes, assuming 8-bit precision for the average intensity values. Alternatively, 100  $16 \times 16$  blocks can be pseudo-randomly located on each frame of the video. For each of these 100 blocks, the first 10 DCT coefficients can be determined. The feature vector then consists of the  $100 \times 10=1000$  DCT coefficients. Many other variations are also possible. In many media applications, the content possesses strong temporal and spatial correlations. If necessary, these correlations can be eliminated or substantially reduced by pre-processing the content with a whitening filter.

A second purpose of the feature extraction process is to acquire a representation that is robust or invariant to possible noise or distortions that a signal might experience. For example, frames of a television broadcast may experience a small amount of jitter, i.e. horizontal and/or vertical translation, or may undergo lossy compression such as MPEG-2. It is advantageous, though not essential, that these and other processes do not adversely affect the extracted vectors.

Each frame's feature vector is then compared with a database of known feature vectors. These known vectors have previously been entered into a content recognition database together with a unique identifier. If a frame's vector matches a known vector, then the commercial is recognized. Of course, there is the risk that the match is incorrect. This type of error is known as a false positive. The false positive rate can be reduced to any desired value, but at the expense of the false negative rate. A false negative occurs when a frame's vector is not matched to the database even though the advertisement is present in the database. There are several reasons why a frame's feature vector may fail to match. First, the recognition system may not be capable of 100% accuracy. Second, the extracted vector will contain noise as a result of the transmission process. This noise may alter the values of a feature vector to the extent that a match is no longer possible. Finally, there is the case where the observed commercial is not yet present in the database. In this case, it is necessary to store the commercial and pass it (e.g., to a person) for identification and subsequent entry in the database.

It is important to realize that the matching of extracted and known vectors is not equivalent to looking up a word in an electronic dictionary. Since the extracted vectors contain noise or distortions, binary search is often not possible. Instead, a statistical comparison is often made between an extracted vector and each stored vector. Common statistical measures include linear correlation and related measures such as correlation coefficient, but other methods can also be used, including clustering techniques. See, e.g., the Duda and

21

Hart reference. These measures provide a statistical measure of the confidence of the match. A threshold can be established, usually based on the required false positive and negative rates, such that if the correlation output exceeds this threshold, then the extracted and known vectors are said to match.

If binary search was possible, then a database containing  $N$  vectors would require at most  $\log(N)$  comparisons. However, in current advertisement monitoring applications there is no discussion of efficient search methods. Thus, a linear search of all  $N$  entries may be performed, perhaps halting the search when the first match is found. On average, this will require  $N/2$  comparisons. If  $N$  is large, this can be computationally expensive. Consider a situation in which one out of 100,000 possible commercials is to be identified. Each 30-second commercial consists of 900 video frames. If all 900 frames are stored in the database, then  $N=90,000,000$ . Even if only every  $10^{\text{th}}$  video frame is stored in the database, its size is still nine million. While databases of this size are now common, they rely of efficient search to access entries, i.e., they do not perform a linear search. A binary search of a 90,000,000-item database requires less than 20 comparisons. In contrast, a linear search will require an average of 45,000,000!

With 9 million entries, if each vector is 1 Kbyte, then the storage requirement is 9 Gigabytes. Disk drives with this capacity are extremely cheap at this time. However, if the database must reside in memory due to real-time requirements, then this still represents a substantial memory requirement by today's standards. One reason that the data may need to be stored in memory is because of the real-time requirements of the database. If 10 channels are being simultaneously monitored within each of 50 geographic areas, then there will be 15,000 queries per second to the content recognition database, assuming each and every frame is analyzed. This query rate is low. However, if a linear search is performed then 675 billion comparisons per second will be required. This is an extremely high computational rate by today's standards. Even if only key frames are analyzed, this is unlikely to reduce the computational rate by more than an order of magnitude.

If an advertisement is not recognized, then typically, the remote monitoring system will compress the video and transmit it back to a central office. Here, the clip is identified and added to the database and the remote recognition sites are subsequently updated. Identification and annotation may be performed manually. However, automatic annotation is also possible using optical character recognition software on each frame of video, speech recognition software, close captioning information and other information sources. As these methods improve in accuracy, it is expected that they will replace manual identification and annotation.

The recognition system described can be considered to be a form of nearest neighbor search in a high dimensional feature space. This problem has been very well studied and is known to be very difficult as the dimensionality of the vectors increases. A number of possible data structures are applicable including kd-trees and vantage point trees. These data structures and associated search algorithms organize a  $N$ -point dataset ( $N=90,000,000$  in out previous example) so that sub-linear time searches can be performed on average. However, worst-case search times can be considerably longer. Recently, Yianilos proposed an excluded middle vantage point forest for nearest neighbor search. See, e.g., the Yianilos reference. This data structure guarantees sub-linear worst-case search times, but where the search is now for a nearest neighbor within a fixed radius,  $\tau$ . The fixed radius search means that if the database contains a vector that is within  $\tau$  of the query,

22

then there is a match. Otherwise, no match is found. In contrast, traditional vantage point trees will always return a nearest neighbor, even if the distance between the neighbor and the query is very large. In these cases, if the distance between the query and the nearest neighbor exceeds a threshold, then they are considered not to match. This is precisely what the excluded middle vantage point forest implicitly does.

Using an excluded middle vantage point forest, will allow accurate real-time recognition of 100,000 broadcasted advertisements. This entails constructing an excluded middle vantage point forest based on feature vectors extracted from say 90,000,000 frames of video. Of course, using some form of pre-filtering that eliminates a large number of redundant frames or frames that are not considered to be good unique identifiers can reduce this number. One such pre-filter would be to only examine the I-frames used when applying MPEG compression. However, this is unlikely to reduce the work identification database (WID) size by more than one order of magnitude. Assuming 10 channels are monitored in each of 50 geographic regions, then the query rate is  $15,000=10 \times 50 \times 30$  queries per second.

#### §4.3.2 Operational Example where Extra-Work Information is Used to Identify the Work

FIG. 8 depicts a satellite television broadcast system **800**, though cable and traditional broadcast modes are also applicable. Block **810** represents audience members (users) watching a TV channel in their home, which also has a connection **812** to the Internet **820**. Other networks are also possible. The satellite broadcasts are also being monitored by one or more television monitoring centers **840a**. These centers **840a** may monitor all or a subset of the television channels being broadcast. They are not restricted to monitoring satellite TV broadcasts but may also monitor cable and traditional terrestrial broadcasts. The primary purpose of these monitoring centers **840a** is to identify the works being broadcast. Of particular interest are television advertisements. However, other works, or portions thereof, may also be identified. Each time a new segment of a work is identified, the monitoring system or systems **840a** update one or more database centers **840b**, informing them of the time, place, channel and identity of the identified segment. The segment may be a complete thirty second commercial or, more likely, updates will occur more frequently, perhaps at a rate of 1 update per second per channel per geographic location. The database center **840b** updates its database so that queries can be efficiently responded to in sub-linear time.

The database centers **840b** can use traditional database technology. In general, the query search initiated by an audience member is not a nearest neighbor search but can be a classical textual search procedure such as a binary search. The nearest neighbor search is appropriate for the monitoring sub-system **840a**. The database centers **840b** are continually updated as each new advertisement, television show or portion thereof is recognized. Standard updating algorithms can be used. However, random new entries to the database are unlikely. Rather, each new entry, or set of entries, denotes a new time segment that is later than all previously inserted items. As such, each new entry can be appended to the end of the database while still maintaining an ordered data structure that is amenable to binary and other efficient search techniques. If two entries have the same time in their time field, items can be sorted based on secondary fields such as the channel and geographic location, as depicted in FIG. 9. Since the number of such entries will be relatively small compared with the entire database, it may be sufficient to simply create

US 8,205,237 B2

23

a linear linked list of such entries, as depicted in FIG. 9. Of course, the size of the database is constantly increasing. As such, it may become necessary to have several levels of storage and caching. Given the envisaged application, most user queries will be for recent entries. Thus, the database may keep the last hours worth of entries in memory. If there is one entry per second for each of 100 channels in 100 geographic locations, this would correspond to  $3600 \times 100 \times 100 = 36,000,000$  entries which is easily accommodated in main memory. Entries that are older than one hour may be stored on disk and entries older than one week may be archived (e.g., backed up on tape) for example. The entries to this database can include time, location and channel information together with a unique identifier that is provided by the monitoring system. Of course, additional fields for each entry are also possible.

When a user query is received, the time, channel and geographic information are used to retrieve the corresponding unique identifier that is then used to access a second database that contains information associated with the identified work.

An entry **1000** in this second database is depicted in FIG. 10, which shows that associated with the unique identifier **1010**, the name of a product **1020**, a product category **1030**, the manufacturer **1040** and the commercial's associated web site **1050**. Many other data fields **1060** are also possible. Such additional fields may include fields that indicate what action should be taken on behalf of the requesting user. Example actions include simply redirecting a request to an associated Web site, or initiating an e-commerce transaction or providing an associated telephone number that may be automatically dialed if the querying device is a cell phone or displaying additional information to the user. This database is likely to be updated much less frequently, perhaps only as often as once or twice a day, as batches of new advertisements are added to the system. Alternatively, it might be updated as each new advertisement is added to the system.

An audience member (user) **810** watching a television commercial for example may react to the advertisement by initiating a query to the database center **840b**. The device whereby the user initiates the query might be a television or set-top-box remote control, or a computer or a wireless PDA or a (WAP-enabled) cell phone or a specialized device. Typically, the query will occur during the airing of the commercial or a shortly thereafter. However, the time between the broadcasting of the advertisement and the time of the associated query is not critical and can, in some instances be much longer. For example, the audience member might bookmark the query information in a device such as a PDA or a specialized device similar to those developed by Xenote for their Itag radio linking. Later, the audience member may transmit the query to the database center **840b**. This might happen hours or even days later.

The query contains information that the database center **840b** uses to identify the work being viewed. This information might include the time and place where the audience member was, together with the channel being viewed. Other identifying information is also possible. The query may also contain additional information that may be used to facilitate the user's transaction and will include the return address of the user. For example, if the user is intending to order a pizza after seeing a Pizza Hut advertisement, the query may also contain personal information including his or her identity, street address and credit card information.

When the database center **840b** receives a query, data in the query is used to identify the work and associated information. A number of possible actions are possible at this point. First, the database center **840b** may simply function as a form of proxy server, mapping the audience member's initial query

24

into a web address associated with the advertisement. In this case, the audience member will be sent to the corresponding Web site. The database center **840b** may also send additional data included in the initial query to this Web site **850** in order to facilitate an e-commerce transaction between the audience member and the advertiser. In some cases, this transaction will not be direct, but may be indirect via a dealer or third party application service provider. Thus, for example, though an advertisement by Ford Motor Company may air nationally, viewers may be directed to different Web sites for Ford dealerships depending on both the audience member's and the dealerships' geographic locations. In other cases, advertisers may have contracted with the database center **840b** to provide e-commerce capabilities. This latter arrangement has the potential to reduce the amount of traffic directed over the public Internet, restricting it, instead to a private network associated with the owner of the database center.

If the audience member (user) is not watching live television but is instead watching a taped and therefore time-shifted copy, then additional processes are needed. For the new generation of digital video recorders, irrespective of the recording media (tape or disk), it is likely to be very easy to include information identifying the location of the recorder, as well as the time and channel recorded. Location information can be provided to the recorder during the setup and installation process, for example. Digital video recorders, such as those currently manufactured by TIVO of Alviso, Calif. or Replay TV of Santa Clara, Calif. have a network connection via telephone, which can then send the query of an audience member to the database center **840b** using the recorded rather than the current information.

In cases where query information has not been recorded, it is still possible to initiate a successful query. However, in this case, it may be necessary to extract the feature vector from the work of interest and send this information to the monitoring center **840a** where the feature vector can be identified. This form of query is computationally more expensive but the relative number of such queries compared to those sent to the database centers **840b** is expected to be small. It should also be noted that the physical separation of the monitoring and database centers, depicted in FIGS. 6 and 7, is not crucial to operation of the system and simply serves to more clearly separate the different functionality present in the overall system configuration.

Although the implementation architectures described above focus on the television media, it is apparent that the present invention is applicable to audio, print and other media.

#### §4.4 CONCLUSIONS

None of the embodiments of the invention require modification to the work or content, i.e., no active signal is embedded. Consequently, there is no change to the production processes. More importantly, from a user perspective, deployment of this system need not suffer from poor initial coverage. Provided the database is sufficiently comprehensive, early adopters will have comprehensive coverage immediately. Thus, there is less risk that the consumer will perceive that the initial performance of the deployed system is poor. Further, the present invention permits statistics to be gathered that measure users' responses to content. This information is expected to be very useful to advertisers and publishers and broadcasters.



US 8,205,237 B2

25

What is claimed is:

1. A computer-implemented method comprising:
  - a) receiving, by a computer system including at least one computer, features that were extracted from a media work by a client device;
  - b) determining, by the computer system, an identification of the media work using the received features extracted from the media work to perform a sub-linear time search of extracted features of identified media works to identify a neighbor; and
  - c) transmitting, by the computer system, information about the identified media work to the client device.
2. The computer-implemented method of claim 1 wherein the media work is an audio work, wherein the features extracted from the work comprise at least one selected from a group consisting of (A) a frequency decomposition of a signal of the audio work, (B) information samples of the audio work, (C) average intensities of sampled windows of the audio work, and (D) information from frequencies of the audio work, and wherein the audio work is one of (A) a broadcast, (B) a digital file, or (C) an MP3 file.
3. The computer-implemented method of claim 1 wherein the information about the identified media work transmitted to the client device includes at least one of (A) a title, or (B) an author.
4. The computer-implemented method of claim 1 further comprising performing an action including at least one of promoting commerce or enhancing interest in the work.
5. Apparatus comprising:
  - a) at least one processor; and
  - b) at least one storage device storing processor-executable instructions which, when executed by the at least one processor, perform a method of
    - 1) receiving features that were extracted from a media work by a client device,
    - 2) determining, by the computer system, an identification of the media work using the features extracted from the media work to perform a sub-linear time search of extracted features of identified media works to identify a neighbor; and
    - 3) transmitting information about the identified media work to the client device.
6. The apparatus of claim 5 wherein the media work is an audio work, wherein the features extracted from the work comprise at least one selected from a group consisting of (A) a frequency decomposition of a signal of the audio work, (B) information samples of the audio work, (C) average intensities of sampled windows of the audio work, and (D) information from frequencies of the audio work, and wherein the audio work is one of (A) a broadcast, (B) a digital file, or (C) an MP3 file.
7. The apparatus of claim 5 wherein the information about the identified media work transmitted to the client device includes at least one of (A) a title, or (B) an author.
8. The apparatus of claim 5 wherein the method further includes performing an action including at least one of promoting commerce or enhancing interest in the work.
9. A computer-implemented method comprising:
  - a) receiving, by a computer system including at least one computer, features that were extracted from media work by a client device;
  - b) determining, by the computer system, an identification of the media work using the received features extracted

26

- from the media work to perform an approximate nearest neighbor search of extracted features of identified media works; and
- c) transmitting, by the computer system, information about the identified media work to the client device.
10. The method of claim 9 wherein the media work is an audio work, wherein the features extracted from the work comprise at least one selected from a group consisting of (A) a frequency decomposition of a signal of the audio work, (B) information samples of the audio work, (C) average intensities of sampled windows of the audio work, and (D) information from frequencies of the audio work, and wherein the audio work one of (A) a broadcast, (B) a digital file, or (C) an MP3 file.
11. The method of claim 9 wherein the information about the identified media work transmitted to the client device includes at least one of (A) a title, or (B) an author.
12. The method of claim 9 further comprising performing an action including at least one of promoting commerce or enhancing interest in the work.
13. Apparatus comprising:
  - a) at least one processor; and
  - b) at least one storage device storing processor-executable instructions which, when executed by the at least one processor, perform a method of
    - 1) receiving features that were extracted from a media work by a client device,
    - 2) determining, by the computer system, an identification of the media work using the received features extracted from the media work to perform an approximate nearest neighbor search of extracted features of identified media works, and
    - 3) transmitting information about the identified media work to the client device.
14. The apparatus of claim 13 wherein the media work is an audio work, wherein the features extracted from the work comprise at least one selected from a group consisting of (A) a frequency decomposition of a signal of the audio work, (B) information samples of the audio work, (C) average intensities of sampled windows of the audio work, and (D) information from frequencies of the audio work, and wherein the audio work is one of (A) a broadcast, (B) a digital file, or (C) an MP3 file.
15. The apparatus of claim 13 information about the identified media work transmitted to the client device includes at least one of (A) a title, or (B) an author.
16. The apparatus of claim 13 wherein the method further includes performing an action including at least one of promoting commerce or enhancing interest in the work.
17. The computer-implemented method of claim 1 wherein the media work is a video signal.
18. The computer-implemented method of claim 17 wherein the video signal is obtained from at least one of (A) a broadcast or (B) a video file format.
19. The computer-implemented method of claim 9 wherein the media work is a video signal.
20. The computer-implemented method of claim 19 wherein the video signal is obtained from at least one of (A) a broadcast or (B) a video file format.
21. The computer-implemented method of claim 1 wherein at least one of the acts of receiving or transmitting is performed via a direct communication between the client device and the computer system.
22. The computer-implemented method of claim 1 wherein at least one of the acts of receiving or transmitting is per-

US 8,205,237 B2

27

formed via an indirect communication between the client device and the computer system.

23. The computer-implemented method of claim 9 wherein at least one of the acts of receiving or transmitting is performed via a direct communication between the client device and the computer system. 5

24. The computer-implemented method of claim 9 wherein at least one of the acts of receiving or transmitting is performed via an indirect communication between the client device and the computer system. 10

25. A computer-implemented method comprising:

a) obtaining, by a computer system including at least one computer, media work extracted features that were extracted from a media work, the media work uploaded from a client device; 15

b) determining, by the computer system, an identification of the media work using the media work extracted features to perform a nonexhaustive search of reference extracted features of reference media works to identify a near neighbor; and

c) determining, by the computer system, an action based on the determined identification of the media work. 20

26. The method of claim 25, wherein the action comprises providing to and/or displaying, at another client device, additional information in association with the media work.

27. The method of claim 26, wherein the additional information is an advertisement.

28. The method of claim 25, wherein the action comprises providing a coupon.

29. The method of claim 25, wherein the action comprises providing a link to a Web site. 25

30. The method of claim 25, wherein the action comprises initiating an e-commerce transaction. 30

28

31. The method of claim 25, wherein the action comprises initiating a telephone call.

32. The method of claim 25, wherein the action comprises logging an event relating to competitive market research data.

33. A computer-implemented method comprising:

a) obtaining, by a computer system including at least one computer, media work extracted features that were extracted from a media work, the media work uploaded from a client device;

b) determining, by the computer system, an identification of the media work using the media work extracted features to perform a sublinear approximate nearest neighbor search of reference extracted features of reference identified media works; and

c) determining, by the computer system, an action based on the determined identification of the media work. 15

34. The method of claim 33, wherein the action comprises providing to and/or displaying, at another client device, additional information in association with the media work.

35. The method of claim 34, wherein the additional information is an advertisement. 20

36. The method of claim 33, wherein the action comprises providing a coupon.

37. The method of claim 33, wherein the action comprises providing a link to a Website. 25

38. The method of claim 33, wherein the action comprises initiating an e-commerce transaction.

39. The method of claim 33, wherein the action comprises initiating a telephone call.

40. The method of claim 33, wherein the action comprises logging an event relating to competitive market research data. 30

\* \* \* \* \*

UNITED STATES PATENT AND TRADEMARK OFFICE  
**CERTIFICATE OF CORRECTION**

PATENT NO. : 8,205,237 B2  
APPLICATION NO. : 11/977202  
DATED : June 19, 2012  
INVENTOR(S) : Cox

Page 1 of 1

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

In the Claims

Column 25, line 64, claim 9, “what” should read --that--.

Column 26, line 14, claim 10, “work one” should read --work is one--.

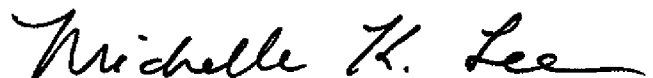
Column 26, line 27, claim 13, “what” should read --that--.

Column 26, line 44, claim 14, “(8)” should read --(B)--.

Column 26, line 46, claim 15, “claim 13 information” should read --claim 13 wherein information--.

Column 26, line 61, claim 20, “B)” should read --(B)--.

Signed and Sealed this  
Sixth Day of May, 2014

A handwritten signature in black ink, reading "Michelle K. Lee". The signature is written in a cursive, flowing style.

Michelle K. Lee  
*Deputy Director of the United States Patent and Trademark Office*